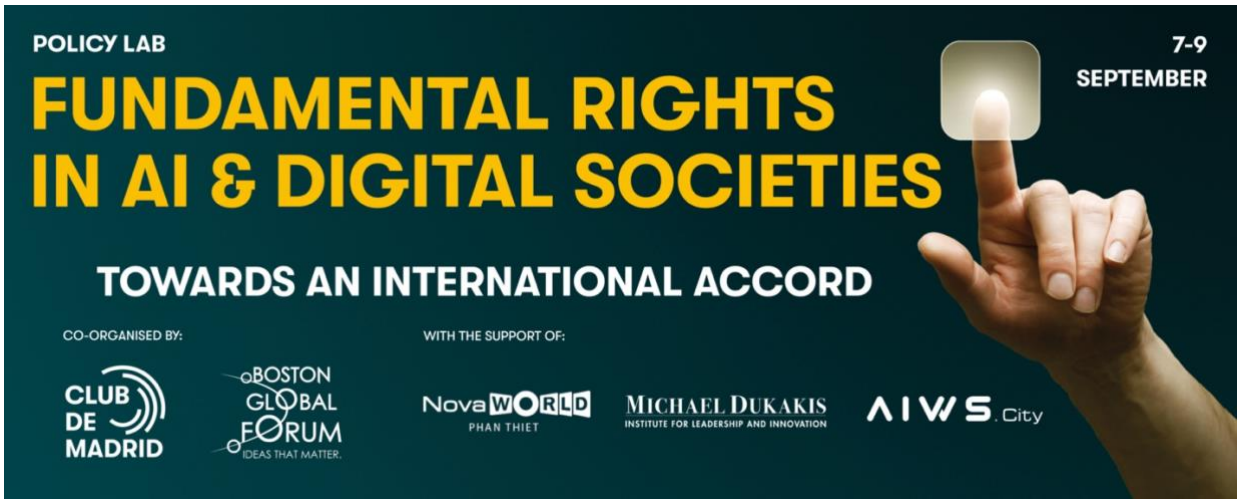


POLICY LAB FUNDAMENTAL RIGHTS IN AI & DIGITAL SOCIETIES: TOWARDS AN INTERNATIONAL ACCORD

Club de Madrid and Boston Global Forum

7-9 September 2021

FINAL REPORT

CLUB DE MADRID

BOSTON GLOBAL FORUM
IDEAS THAT MATTER.

MICE PHAN THIET

MICHAEL DUKAKIS
INSTITUTE FOR LEADERSHIP AND INNOVATION

AIWS City

POLICY LAB
FUNDAMENTAL RIGHTS IN AI & DIGITAL SOCIETIES:
TOWARDS AN INTERNATIONAL ACCORD

Final Report

1. Background

Since its foundation, **Club de Madrid (CdM)** has worked for the strengthening of democracy through the experienced voice and agency of its Members, democratic former Presidents and Prime Ministers from around the globe. CdM has also been working to promote a strengthened multilateral system that provides global solutions to global challenges through inclusive cooperation that leaves no one behind.

Digital transformation is one of the global challenges on which CdM and the **Boston Global Forum (BGF)** have been working for a few years. With social norms potentially compromised by the velocity of digitalization, artificial intelligence and social media, and gaps in the rules-based multilateral trading system generating trade and technology tensions between States, both organizations found that a foundational, norm-setting initiative was imperative to ensure **we manage digitalization before it manages us.**

The digitalization process is rapidly altering the panorama for fundamental rights around the world. Digital technologies and artificial intelligence (AI) create new opportunities for the enjoyment of human rights, as well as new threats to their protection. They also bring about new concepts of rights and obligations, directly arising from the relationship between citizens and technology.

Governing this transformation, with the protection of fundamental rights as a central objective, requires a new approach to public policy, in line with the systemic changes occurring in society, the economy and international relations. While there have been attempts to articulate principles for an approach to the governance of digital societies based on the protection of fundamental rights, **the international community has yet failed to define and adopt a common framework.**

With the invaluable support of **NovaWorld Phan Thiet** and partnership of the BGF, CdM seeks to contribute to **global consensus-building around a rights-based agenda for the governance of digital societies.** This initiative will feed into the work of the AI World Society, an innovation network created by BGF to put forward new principles for governance in the age of AI. In partnership with UN Academic Impact, BGF and the AI World Society have launched the UN Centennial Initiative. Its *core is the e-book "Remaking the World – Toward an Age of Global Enlightenment"*, for which they will continue to develop and promote models for a *Social Contract for the AI Age* and an *AI International Accord*, using the AIWS City as a sandbox to put their ideas into practice.

Over 2021-2025, CdM will feed into these efforts by convening annual policy discussions of different sizes and scopes, combining general reflections on fundamental rights in the age of AI together with more issues-based policy discussions around specific areas where digital technologies and AI are being deployed.

2. Policy Lab on Fundamental Rights in AI & Digital Societies: Towards an International Accord

In this context, CdM in partnership with BGF organized a Policy Lab on *Fundamental Rights in AI & Digital Societies: Towards an International Accord* in September 2021. Multistakeholder discussions aimed to build consensus around a rights-based agenda for the global governance of AI and digital societies, focusing on the following topics:

- Opportunities and threats for fundamental rights in AI & digital societies.
- Transatlantic approaches to protect fundamental rights in AI & digital spaces.

- The elements & processes for an international legal framework to protect fundamental rights in AI & digital spaces
- The Concept, Principle and Ecosystem for Digital and AI Society – "Remaking the World, Toward an Age of Global Enlightenment".
- The Global Alliance for Digital Governance
- A framework for a global law and accord on AI and digital tools

The Policy Lab followed the work initiated by both organizations on the implications of digital technologies and AI for democracy, global cooperation and multilateralism, with a particular focus on the Transatlantic space, and based on the progress of the AIWS Social Contract 2020 (Social Contract for the age of AI) and AIWS Innovation Network (AIWS.net).

The anchor of the discussions was a vision for a human-centered digital age. Policymakers around the world and at all levels of government are becoming more convinced of the need to ensure that digital technologies and AI serve the people. CdM and BGF partnered to narrow this gap between the digital and policy-making through a multi-stakeholder discussion that harnessed the political experience and insights of ten CdM Members, current political leaders, representatives from business, academia and other high-level experts.

3. Activities and methodology

The Policy Lab was held as a **three-day online event** from **7 to 9 September 2021.**

During these three days, participants reflected on the intersection of digital technologies, fundamental rights and global governance, analyzing the challenges and opportunities for the global community to agree on a basic set of principles to protect and promote fundamental rights in AI and digital societies.

The Policy Lab was divided into **keynotes**, and **plenary discussions** in a multi-stakeholder setting and as an integrated event production. This meant that the speakers and participants – regardless of their location – were virtually connected by an online platform.

3.1. Sub-Committees

During the months leading up to the Policy Lab, three **Sub-Committees** analyzed the **opportunities and threats for fundamental rights in AI and digital societies**; **Transatlantic approaches to protect fundamental rights in AI and digital spaces, and the elements and process for an international legal framework to protect fundamental rights in AI and digital spaces,** with the Concept and Ecosystem for Digital and AI Society – "Remaking the World, Toward an Age of Global Enlightenment", and the Global Alliance for Digital Governance as cross-cutting themes.

The Sub-Committees were multidisciplinary teams composed of CdM Members – democratic former Heads of State and Government - as well as experts from governments, international organizations, civil society, academia, the private sector, and decision-makers.

Based on the analysis of the status and current existing frameworks in the three key issues abovementioned, Sub-Committees prepared **briefing issues papers** in each field, and formulated preliminary, actionable policy proposals that fueled the *Policy Lab's Plenary I: Setting the context for a rights-based international accord*. These Briefing Issues Papers are included as annexes to this Report.

Participants in each Sub-Committee were:

Sub-Committee 1: Opportunities and threats for fundamental rights in AI and digital societies

Sub-Committee leader: **Paul Twomey**, Fellow and Core Theme Leader at the Global Solutions Initiative; Distinguished Fellow at the Centre for International Governance Innovation; Commissioner of the Global Commission for Internet Governance and former CEO of the Internet Corporation for Assigned Names and Numbers (ICANN)

Sub-Committee members:

- **Hanna Suchocka**, Member of Club de Madrid, Prime Minister of Poland (1992-1993)
- **Sean Cleary**, Advisor of Club de Madrid, Executive Vice-Chair of the FutureWorld Foundation, Member of the Carnegie Council's Artificial Intelligence & Equality Initiative's Board of Advisors
- **Thomas Patterson**, Research Director of The Michael Dukakis Institute for Leadership and Innovation, Professor of Government and the Press of Harvard Kennedy School
- **Gregor Strojin**, Chair of the Council of Europe's Committee on Artificial Intelligence (CAHAI)
- **Jutta Treviranus**, Professor and Director of Inclusive Design Institute at the Ontario College of Art & Design (OCAD)

Sub-Committee 2: Transatlantic approaches to protect fundamental rights in AI and digital spaces

Sub-Committee leaders:

- **Jerry Jones**, Advisor of Club de Madrid, Executive Vice-President, Ethics and Legal Officer, Live Ramp
- **Véronique Choquette**, Senior Policy & Programme Development Advisor at Club de Madrid

Sub-Committee members:

- **Zlatko Lagumdžija,** Member of Club de Madrid, Prime Minister of Bosnia and Herzegovina (2001-2002)
- **Irene Braam,** Executive Director, Bertelsmann Foundation North America
- **Manuel Muñiz,** Provost of IE University, Dean of IE School of Global and Public Affairs, Former Secretary of State for Global Spain at the Spanish Foreign Ministry

Sub-Committee 3: The elements and process for an international legal framework to protect fundamental rights in AI and digital spaces, towards an AI international accord.

Sub-Committee leader: **Paul Nemitz,** Principal Advisor in the Directorate General for Justice and Consumers of the European Commission

Sub-Committee members:

- **Danilo Türk,** President of Club de Madrid President of Slovenia (2007-2012)
- **Nazli Choucri,** Boston Global Forum Board Member and Professor of Political Science at the Massachusetts Institute of Technology (MIT)
- **Eva Kaili,** Member of EU Parliament

3.2. Plenary sessions

Danilo Türk, President of Club de Madrid, President of Slovenia (2007-2012) and Governor Michael Dukakis, Chairman of the Michael Dukakis Institute for Leadership and Innovation & Co-Founder of the Boston Global Forum, opened the Policy Lab.

Following the opening session, Robin Kelly, Member of the U.S. House of Representatives from Illinois's 2nd district, gave the **Keynote Speech** focused on a **human-centered digital age.**

The Policy Lab had **six plenary sessions** with the following titles and objectives:

Plenary I: Setting the context for a rights-based international accord

Plenary I aimed to define the current context and opportunities for change, proposing solutions and recommendations to governing digital transformation, with the protection of fundamental rights as a central objective.

The three Sub-Committees' insights and the ensuing panel referred to the three main topics addressed in the Sub-Committees:

- Opportunities and threats for fundamental rights in AI & digital societies
- Transatlantic approaches to protect fundamental rights in AI & digital spaces
- Elements & processes for an international legal framework to protect fundamental rights in AI & digital spaces

Participants in Plenary I were as follows:

Sub-committees' Insights:

- Paul Twomey, Fellow and Core Theme Leader at the Global Solutions Initiative; Distinguished Fellow at the Centre for International Governance Innovation; Commissioner of the Global Commission for Internet Governance and former CEO of the Internet Corporation for Assigned Names and Numbers (ICANN)
- Jerry Jones, Advisor of CdM, Executive Vice-President, Ethics and Legal Officer, Live Ramp
- Paul Nemitz, Principal Advisor in the Directorate General for Justice and Consumers of the European Commission

Facilitator: Véronique Choquette, Senior Policy & Programme Development Advisor at Club de Madrid

Panel discussion

- Hanna Suchocka, Member of Club de Madrid, Prime Minister of Poland (1992-1993)
- Kim Campbell, Member of Club de Madrid, Prime Minister of Canada (1993)
- Paul Twomey, Fellow and Core Theme Leader at the Global Solutions Initiative; Distinguished Fellow at the Centre for International Governance Innovation;

Commissioner of the Global Commission for Internet Governance and former CEO of ICANN

- Jerry Jones, Advisor of Club de Madrid, Executive Vice-President, Ethics and Legal Officer, Live Ramp
- Paul Nemitz, Principal Advisor in the Directorate General for Justice and Consumers at the European Commission

Plenary II: Regional perspectives and models for the governance of AI and digital societies

Plenary II analysed the most relevant points of the regional proposals and other models that could serve as a basis to advance in the elaboration of a rights-based international agreement on digital technologies and AI. In addition to the dominant and much-discussed EU and US approaches to AI and digital technologies, this session also expanded the reflections to other regions of the world, looking at how different parts of the geopolitical puzzle could play out for the democratic rights-based governance of AI and digital societies.

Participants in Plenary II were as follows:

Panel discussion

- Kevin Rudd, Member of Club de Madrid, Prime Minister of Australia (2007-2010, 2013), (video-message)
- Iveta Radičová, Member of Club de Madrid, Prime Minister of Slovakia (2010-2012)
- Irene Braam, Executive Director, Bertelsmann Foundation North America
- Nanjira Sambuli, Fellow in the Technology and International Affairs Program at The Carnegie Endowment for International Peace
- David Bray, Director of GeoTech Center and GeoTech Commission, Atlantic Council
- Manuel Muñiz, Provost of IE University, Dean of IE School of Global and Public Affairs, Former Secretary of State for Global Spain at the Spanish Foreign Ministry
- Marcelo Cabrol, Manager of the Inter-American Development Bank (IDB) Social Sector

Facilitator: Ramu Damodaran, Co-Chair of the United Nations Centennial Initiative

Plenary III: Current frameworks and agreements for AI International Accord (AIIA)

Starting from analysing strategies at the local, national and multilateral level, Plenary III discussed the key points that would allow policy-makers and decision-makers to put technology at the service of people, assessing the essential elements of current strategies to advance towards an AI international accord. The Framework for AIIA that appears in Annex V, proposes a set of measures addressed to actors and entities for immediate review, assessment, refinement, and adoption by the international community.

Participants in Plenary III were as follows:

Lead speaker: Cameron Kerry, Former Acting Secretary of the U.S. Department of Commerce

Panel discussion

- Danilo Türk, President of Club de Madrid President of Slovenia (2007-2012)
- Doris Leuthard, Member of Club de Madrid, President of the Swiss Confederation (2010 and 2017)
- Nazli Choucri, Boston Global Forum Board Member and Professor of Political Science at the Massachusetts Institute of Technology (MIT)
- Gabriela Ramos, Assistant Director-General for the Social and Human Sciences of UNESCO
- Jutta Treviranus, Professor and Director of Inclusive Design Institute at the Ontario College of Art & Design (OCAD)
- Carlos Santiso, Director of Digital Innovation in Government Directorate at Development Bank of Latin America – CAF

Facilitator: Véronique Choquette, Senior Policy & Programme Development Advisor at Club de Madrid

Plenary IV: A Global Alliance for Digital Governance

The rapid deployment and decentralisation of new technologies and AI beyond the control of States and digital interdependence in this globalised world have led to a sweeping set of interrelated challenges that requires a division of responsibility and the articulation of collective responses in all sectors and levels.

In this context, Plenary IV analysed ways to coordinate resources between governments, international organisations, corporations, think tanks, civil society, and champions on AI and digital governance to build a peaceful, secure, prosperous, and human-centred world in the age of AI and digitalisation, aimed at moving forward to a Global Alliance for Digital Governance.

Participants in Plenary IV were as follows:

Lead speaker: Eva Kaili, Member of EU Parliament

- Aminata Touré, Member of Club de Madrid, Prime Minister of Senegal (2013-2014)
- Nguyen Anh Tuan, CEO of the Boston Global Forum
- Vilas Dhar, President and Trustee of the Patrick J. McGovern Foundation
- Donara Barojan, Co-founder and CEO of Polltix

Facilitator: John Quelch, Dean of the University of Miami Herbert Business School

Plenary V: The United Nations Centennial Initiative: The Practice of Fundamental Rights in AI & Digital Societies

Plenary V was based on the United Nations Centennial initiative, launched by the BGF and the United Nations Academic Impact (UNAI) in 2019 as the United Nations planned to mark its 75th anniversary the following year. The initiative brought into its fold some of the finest minds of our times as they sought to anticipate the world, and the United Nations, in 2045, the year of the organization's centennial. The core concepts of the initiative are reflected in the book "Remaking the World: Toward an Age of Global Enlightenment"; these include the idea of a social contract for the Artificial Intelligence (AI) age, a framework for an AI international accord, an ecosystem for the "AI World Society" (AIWS) and a community innovation economy. Some of these ideas

have already begun to be put into practice, including a Global Alliance for Digital Governance and the evolution of an AIWS City being developed by NovaWorld in Phan Thiet, Viet Nam.

Following the work started by BGF and UNAI, Plenary V identified ideas that could support the UN effort to maintain international peace and security, examining pressing issues surrounding technology on these matters, including AI, cybersecurity, diplomacy, and other concerns that affect the defence and promotion of fundamental rights in the digital sphere.

Participants in Plenary V were as follows:

Lead speaker: Ramu Damodaran, Co-Chair of the United Nations Centennial Initiative

Panel discussion:

- Vaira Vike-Freiberga, Member of Club de Madrid, President of Latvia (1999-2007)
- Kyriakos Pierrakakis, Minister of State and Digital Governance of Greece, Chair of the Global Strategy Group, OECD
- Thomas Patterson, Research Director of the Michael Dukakis Institute for Leadership and Innovation, Professor of Government and the Press of Harvard Kennedy School
- Sean Cleary, Advisor of Club de Madrid, Executive vice-chair of the FutureWorld Foundation, Member of the Carnegie Council's Artificial Intelligence & Equality Initiative's Board of Advisors
- Tran Dinh Thien, Professor and Senior Advisor to Vietnamese Prime Minister

Facilitator: David Silbersweig, Chairman, Department of Psychiatry and Co-Director for Institute for the Neurosciences, Brigham and Women's Hospital, Harvard Professor

Plenary VI: Building safer, equitable and trustworthy AI and digital societies: The AI International Accord (AIIA)

After examining and analysing the existing principles, values and standards related to digital technologies and AI, Plenary VI analysed whether current conditions allow for progress towards a rights-based international framework that fits the AI and digital era and which would be the more effective way to that end.

Participants in Plenary VI were as follows:

Lead speaker: Alex 'Sandy' Pentland, Director, MIT Connection Science and Human Dynamics labs, "Making the New Social Contract Work"

- Zlatko Lagumdžija, Member of Club de Madrid, Prime Minister of Bosnia and Herzegovina (2001-2002)
- Esko Aho, Member of Club de Madrid, Prime Minister of Finland (1991-1995)
- Gregor Strojin, Chair of the Council of Europe's Committee on Artificial Intelligence (CAHAI)
- Karine Caunes, Global Program Director, Center for AI and Digital Policy (CAIDP)

Facilitator: William Hoffman, Head of Data-Driven Development, World Economic Forum

Finally, Nguyen Anh Tuan, CEO of the Boston Global Forum, Director of the Michael Dukakis Institute for Leadership and Innovation and María Elena Agüero, Secretary General of Club de Madrid, gave the closing words and next steps.

4. Key conclusions

- There is no doubt that digital technologies, and AI, in particular, have, for better or for worse, generated a revolution for fundamental rights. Building an international agreement on digital governance has complexities and the global policy and geopolitical environment plays a key role in facilitating or limiting the construction of this agreement.

- Common democratic values such as **respect and promotion of human rights, and the rule of law** are crucial to underpinning digital policy as an essential starting point to move towards that agreement.

- Challenges such as AI and data governance that **domestic frameworks** cannot address alone are crucial points on which we must focus. From there, we can start with small but important steps to build a culture of agreement on digital issues – with a premium on the **Transatlantic space**, that has the advantage of **shared values.**

- In a field where so much is yet to come, we are convinced that **international cooperation for Artificial Intelligence and digital technologies** is an opportunity to write the **rules together.** The Framework for AI International Accord, a part of the e-book "Remaking the World – Toward an Age of Global Enlightenment", presented at this Policy Lab is a significant start for this goal.

- We need some internationally agreed fundamental rules or norms to guide the development of technologies; we cannot anticipate to protect rights we do not fully comprehend; and the efforts that already exist are essential to continue working on the objective that gathered us these three days. It will be a challenging process, because of the variety of values and approaches that are emerging in different parts of the world, but there is common ground to be found. And to that end, **making principles operational** and integrating a **variety of stakeholders** representing countries and communities in all their diversity, including inter-generational differences is needed.

- Many of the issues discussed intersect with the **crucial work the UN is both doing and planning to do**, under the leadership of **Secretary-General Guterres**, to maintain **international peace and security, and support the achievement of the SDGs**. AI, cybersecurity, diplomacy, and development – not least social development – all relate to defense and promotion of fundamental rights in the digital sphere. It is our aim that our recommendations, the United Nations Centennial Initiative, and the book **"Remaking the World – Toward an Age of Global Enlightenment"** support *'Our Common Agenda'* and, particularly, the **Global Digital Compact** proposal.

- There is no lack of goodwill and effort to build an AI framework on which different actors – governments, local governments, and non-government actors can agree. The UNESCO Recommendation on the Ethics of Artificial Intelligence is a promising step in the right direction.

- We have also established a **Global Alliance for Digital Governance** that includes relevant stakeholders -governments, private sector, academia, civil society, international

organizations- to reduce the digital field's development gaps and bring communities together, thus contributing to the United Nations Centennial Initiative.

- We agreed on the need for a **new social contract** that takes digital transformation into account. To build a social contract suited for the digital age, going beyond traditional allies and reach out to those who think differently is crucial. The Social Contract for the AI Age is a recognized tool and will be fundamental for the Age of Global Enlightenment.

- Throughout this process of reflection, **trust** is essential and to obtain that we would need to build on security, privacy, reliability and fairness as crucial pillars that will promote digital technologies as a tool to serve inclusive societies.

- Protecting access to **information, education** and **digital literacy** and finding a balance between freedom of speech and the imperative to have a **common truth** will allow progress on drafting common rules on AI. In this regard, the AIWS City will be a practical model for addressing this issue.

- It is tough to craft legislation and rules for technologies that are not yet being used, so we need a **risk-based approach** to digital governance. In the case of AI, this approach will help to elaborate some of the requirements for its design, development and application phases.

- ***Ex ante* and *ex-post* regulation are not incompatible**. We need both to better govern digital. *Ex ante* regulation will allow institutions to provide guardrails for rights, including data rights, in the deployment of AI systems. *Ex post* regulation will allow AI systems to be audited. In this regard, we agreed **accountability** is a fundamental consideration in the deployment of AI technologies. We need to be able to explain how AI systems reach the decisions they reach and will allow us to work to stop the dynamics of discrimination, exclusion and inequalities that are being replicated and amplified by AI technologies. The **Global Alliance for Digital Governance** can be a significant movement for this mission.

- The **Community Innovation Economy** concept was introduced during the Policy Lab as a tool that empowers citizens to create value for themselves, for others, and for society through the application of AI, digital, block chain, and data science technologies. It is a sharing ecosystem that rewards both the creators and users of these technologies, as well as an ecosystem that encourages everyone to innovate.

- Despite the existing gaps in the regulation of digital technologies and their use, they have been fundamental **tools of resilience** during the COVID-19 pandemic and we must not forget their benefits.

- Finally, we would like to mention that many of the discussions of the six Plenaries highlighted the significant contributions of the e-book, "Remaking the World - Toward an Age of Global Enlightenment", published by the UN Centennial Initiative and the Boston Global Forum.

5. Next Steps

- This Policy Lab is a starting point of a 5-year initiative in which we propose to steer global conversations on fundamental rights in digital societies, to build bridges between

countries, regions and communities of practices, and identify a path to consensus around a rights-based agenda for the governance of AI and digital technologies.

- In Club de Madrid **Annual Policy Dialogue** which this year will focus on **'Rethinking Democracy: A Global Agenda for Democratic Renewal'** and will take place from the **27 to 29 October**, we will again look at the challenge of digital technologies and artificial intelligence, but focusing on the **new information ecosystem** increasingly driven by digital platforms and how to build democratic approaches to reconcile truth, trust and freedom in this ecosystem.

- Policy Lab conclusions, will feed into our October Policy Dialogue when we will again be partnering among others with the Boston Global Forum, its renowned scholars and the Global Alliance for Digital Governance, which will serve to coordinate with distinguished leaders, strategists, thinkers, and innovators, the creation of a Global Law and Accord on AI and Digital, and contribute concepts for a mechanism that has enough power to enforce them.

- Recommendations will be disseminated within both **national governments** and **multilateral decision-makers** and, in this way, feed into the most relevant political processes related to the governance of the digital transformation. They will also be incorporated into the 5-year plan of the CdM-BGF project framework.

POLICY LAB
FUNDAMENTAL RIGHTS IN AI & DIGITAL SOCIETIES: TOWARDS AN INTERNATIONAL ACCORD

AGENDA

DAY 1 – September 7, 15:00-17:30 CEST

15:00 – 15:05 **Introduction to the Policy Lab on Fundamental Rights in AI & Digital Societies: Towards an International Accord**

Master of Ceremonies: Véronique Choquette, Senior Policy & Programme Development Advisor at Club de Madrid

15:05 – 15:15 **Opening Session**

- Danilo Türk, President of Club de Madrid, President of Slovenia (2007-2012)
- Governor Michael Dukakis, Chairman of the Michael Dukakis Institute for Leadership and Innovation & Co-Founder of the Boston Global Forum

15:15 – 15:25 **Keynote Speech: 'A human-centered digital age'**

- Robin Kelly, Member of the U.S. House of Representatives from Illinois's 2nd district

15:25 – 16:25 **Plenary I: Setting the context for a rights-based international accord**

Sub-committees' Insights:

- Paul Twomey, Fellow and Core Theme Leader at the Global Solutions Initiative; Distinguished Fellow at the Centre for International Governance Innovation; Commissioner of the Global Commission for Internet Governance and former CEO of the Internet Corporation for Assigned Names and Numbers (ICANN)
- Jerry Jones, Advisor of CdM, Executive Vice-President, Ethics and Legal Officer, Live Ramp
- Paul Nemitz, Principal Advisor in the Directorate General for Justice and Consumers of the European Commission

Facilitator: Véronique Choquette, Senior Policy & Programme Development Advisor at Club de Madrid

Panel discussion

- Hanna Suchocka, Member of Club de Madrid, Prime Minister of Poland (1992-1993)
- Kim Campbell, Member of Club de Madrid, Prime Minister of Canada (1993)
- Paul Twomey, Fellow and Core Theme Leader at the Global Solutions Initiative; Distinguished Fellow at the Centre for International Governance Innovation; Commissioner of the Global Commission for Internet Governance and former CEO of ICANN

- Jerry Jones, Advisor of Club de Madrid, Executive Vice-President, Ethics and Legal Officer, Live Ramp
- Paul Nemitz, Principal Advisor in the Directorate General for Justice and Consumers at the European Commission

16:25 – 17:25 **Plenary II: Regional perspectives and models for the governance of AI and digital societies**

Facilitator: Ramu Damodaran, Co-Chair of the United Nations Centennial Initiative

Panel discussion

- Kevin Rudd, Member of Club de Madrid, Prime Minister of Australia (2007-2010, 2013), (video-message)
- Iveta Radičová, Member of Club de Madrid, Prime Minister of Slovakia (2010-2012)
- Irene Braam, Executive Director, Bertelsmann Foundation North America
- Nanjira Sambuli, Fellow in the Technology and International Affairs Program at The Carnegie Endowment for International Peace
- David Bray, Director of GeoTech Center and GeoTech Commission, Atlantic Council
- Manuel Muñiz, Provost of IE University, Dean of IE School of Global and Public Affairs, Former Secretary of State for Global Spain at the Spanish Foreign Ministry
- Marcelo Cabrol, Manager of the Inter-American Development Bank (IDB) Social Sector

17:25 – 17:30 **Conclusions Day 1 and closing**

DAY 2 – September 8, 15:00-17:30 CEST

15:00 – 15:05 **Introduction**

Master of Ceremonies: Véronique Choquette, Senior Policy & Programme Development Advisor at Club de Madrid

15:05 – 16:05 **Plenary III: Current frameworks and agreements for an AI International Accord**

Facilitator: Véronique Choquette, Senior Policy & Programme Development Advisor at Club de Madrid

Lead speaker: Cameron Kerry, Former Acting Secretary of the U.S. Department of Commerce

Panel discussion

- Danilo Türk, President of Club de Madrid President of Slovenia (2007-2012)
- Doris Leuthard, Member of Club de Madrid, President of the Swiss Confederation (2010 and 2017)
- Nazli Choucri, Boston Global Forum Board Member and Professor of Political Science at the Massachusetts Institute of Technology (MIT)
- Gabriela Ramos, Assistant Director-General for the Social and Human Sciences of UNESCO

- Jutta Treviranus, Professor and Director of Inclusive Design Institute at the Ontario College of Art & Design (OCAD)
- Carlos Santiso, Director of Digital Innovation in Government Directorate at Development Bank of Latin America – CAF

16:05 – 17:05 **Plenary IV: A Global Alliance for Digital Governance**

Facilitator: John Quelch, Dean of the University of Miami Herbert Business School

Lead speaker: Eva Kaili, Member of EU Parliament

- Aminata Touré, Member of Club de Madrid, Prime Minister of Senegal (2013-2014)
- Nguyen Anh Tuan, CEO of the Boston Global Forum
- Vilas Dhar, President and Trustee of the Patrick J. McGovern Foundation
- Donara Barojan, Co-founder and CEO of Polltix

17:05 – 17:10 **Conclusions Day 2**

17:10 - 17:30 **Networking session –** Open the floor for participants to present their work

- Nova World Phan Thiet presents

DAY 3 – September 9, 15:00-17:30 CEST

15:00 – 15:05 **Introduction**

Master of Ceremonies: Véronique Choquette, Senior Policy & Programme Development Advisor at Club de Madrid

15:05 – 16:05 **Plenary V: The United Nations Centennial Initiative: The Practice of Fundamental Rights in AI & Digital Societies**

Facilitator: David Silbersweig, Chairman, Department of Psychiatry and Co-Director for Institute for the Neurosciences, Brigham and Women's Hospital, Harvard Professor

Lead speaker: Ramu Damodaran, Co-Chair of the United Nations Centennial Initiative

Panel discussion:

- Vaira Vike-Freiberga, Member of Club de Madrid, President of Latvia (1999-2007)
- Kyriakos Pierrakakis, Minister of State and Digital Governance of Greece, Chair of the Global Strategy Group, OECD
- Thomas Patterson, Research Director of The Michael Dukakis Institute for Leadership and Innovation, Professor of Government and the Press of Harvard Kennedy School
- Sean Cleary, Advisor of Club de Madrid, Executive vice-chair of the FutureWorld Foundation, Member of the Carnegie Council's Artificial Intelligence & Equality Initiative's Board of Advisors
- Tran Dinh Thien, Professor and Senior Advisor to Vietnamese Prime Minister

16:05 – 17:05 **Plenary VI: Building safer, equitable and trustworthy AI and digital societies: The AI International Accord (AIIA)**

Facilitator: William Hoffman, Head of Data-Driven Development, World Economic Forum

Lead speaker: Alex 'Sandy' Pentland, Director, MIT Connection Science and Human Dynamics labs, "Making the New Social Contract Work"

- Zlatko Lagumdžija, Member of Club de Madrid, Prime Minister of Bosnia and Herzegovina (2001-2002)
- Esko Aho, Member of Club de Madrid, Prime Minister of Finland (1991-1995)
- Gregor Strojin, Chair of the Council of Europe's Committee on Artificial Intelligence (CAHAI)
- Karine Caunes, Global Program Director, Center for AI and Digital Policy (CAIDP)

17:05 – 17:20 **Closing words and next steps**

- Nguyen Anh Tuan, CEO of the Boston Global Forum (BGF), Director of the Michael Dukakis Institute for Leadership and Innovation
- María Elena Agüero, Secretary General of Club de Madrid (CdM)

Club de Madrid/Boston Global Forum

POLICY LAB
FUNDAMENTAL RIGHTS IN AI DIGITAL SOCIETIES: TOWARDS AN INTERNATIONAL ACCORD

Issues paper from sub-committee 1:Opportunities and threat for fundamental rights in AI & digital societies.

Dr Paul Twomey
Sub-Committee Leader

Artificial Intelligence (AI) is reshaping human experience in ways not visible to, nor fully apprehended by, the vast majority of the world's population. The explosion of AI is having a notable impact on our present rights and future opportunities, determining the decision-making processes that affect all in today's society.

Enormous technological change is occurring. It promises great benefits and poses insidious risks. The proportion of risks to benefits will become apparent, depending on the pioneers and creators of this technology, and, in particular, on the clarity of their and the political classes' vision of the common good.

This issue paper from the Sub-Committee commences with a discussion of the issues posed by how AI (and its interrelated Big Data) is used in the work place, the market for consumer and citizen behavior, and in the military.

Then the paper turns to the questions of competition issues and impact on human rights. Further some principles for government responses are outlined.

Fourthly multilateral governmental responses to date are briefly sketched.

The above issues were discussed at a videoconference among some of the members of the Subcommittee. The final section of this paper indicates the suggestions for the Policy Lab from this discussion.

The Issues

The use of automated decision making informed by algorithms is penetrating the modern workplace, and broader society, at a rapid rate. In ways not visible to, nor fully apprehended by, the vast majority of the population, algorithms are determining our present rights and future opportunities. To consider just one

aspect of everyday life, automobile transportation, these algorithms help us drive our cars, determine whether we can get a loan to buy them, decide which roads should be repaired, identify if we have broken the rules of the road and even determine whether we should be imprisoned if we have (see Angwin et al. 2016).

Benefits

Big data and AI can provide many benefits. They can assemble and consider more data points than humans can incorporate and often provide less biased or clearer outcomes than humans making decisions.

Examples include the prevention of medical errors to increasing productivity and reducing risks in the workplace. Even in the explicitly human function of the human resources department, machine learning can improve job descriptions and provide more "blind" recruitment processes, which can both increase the pool of qualified candidates and boost recruitment of non-conventional applicants.[1] Written well, algorithms can be more impartial and pick up patterns people may miss, in this and other applications.
Many commentators point to the productivity benefits of AI. For instance, analysis by Accenture of 12 developed economies indicates that AI could double annual economic growth rates in 2035: "The impact of AI technologies on business is projected to increase labor productivity by up to 40 percent and enable people to make more efficient use of their time" (Purdy and Daugherty 2016). The World Bank is exploring the benefits of AI for development and in uses from predicting migration patterns to reducing poverty.[2] Others identify farming, resource provision and health care as sectors in the developing economies that will benefit greatly from the application of AI (see Ovenden 2016).

Impact on Employment

Much has been made of the impact of AI and related robotics on jobs, especially since Carl Benedikt Frey and Michael A. Osborne's 2013 paper estimating that 47 percent of jobs in the United States were "at risk" of being automated in the next 20 years. Debate has ensued on the exact nature of this impact: the full or partial erosion of existing job tasks, the impacts across sectors and across developed, emerging and developing economies. Forecasting such effects is inherently difficult. But a recent summary from the McKinsey Global Institute reflects a midway analysis.

> Automation technologies including artificial intelligence and robotics will generate significant benefits for users, businesses, and economies, lifting productivity and economic growth. The extent to which these technologies displace workers will depend on the pace of their development and adoption, economic growth, and growth in demand for work. Even as it causes declines in some occupations, automation will change many more — 60 percent of occupations have at least 30 percent of constituent work activities that could be automated. It will also create new occupations that do not exist today, much as technologies of the past have done…
> Our scenarios across 46 countries suggest that between almost zero and one-third of work activities could be displaced by 2030, with a midpoint of 15 percent. The proportion varies widely across countries, with advanced economies more affected by automation than developing ones, reflecting higher wage rates and thus economic incentives to automate….

[1] See firms like Textio (https://www.textio.com/) and Pymetrics (https://www.pymetrics.com).
[2] See www.measuredev.org/.

> Even if there is enough work to ensure full employment by 2030, major transitions lie ahead that could match or even exceed the scale of historical shifts out of agriculture and manufacturing. Our scenarios suggest that by 2030, 75 million to 375 million workers (3 to 14 percent of the global workforce) will need to switch occupational categories. Moreover, all workers will need to adapt, as their occupations evolve alongside increasingly capable machines. (Manyika et al. 2017, vi)

Whatever the specifics, the results are clearly going to be very significant for G20 economies and their citizens. And, if the rate of adoption continues to outpace previous major technological adoptions,[3] the scale of social dislocation is likely to be greater — which provides even more reason for the G20 to work now on a framework for AI adoption.

Risk of Bias

Code is written by humans and its complexity can accentuate the flaws humans naturally bring to any task.

Bias in the writing of algorithms, as a product of human endeavour, is inevitable, and can have chilling effects on individual rights, choices and the application of worker and consumer protections. Algorithms incorporate built-in values and serve business models, which may lead to unintended biases, discrimination or economic harm.[4] Compounding this problem is the fact that algorithms are often written by relatively inexperienced programmers who may not have a correct picture of the entire application or a broad experience of a complex world. The dependency of the workplace on algorithms imparts tremendous power to those who write them. These programmers may not even be aware of this power or the potential harm that an incorrectly coded algorithm could do. Researchers have discovered bias in the algorithms for systems used for university admissions, human resources, credit ratings, banking, child support systems, social security systems and more. Because the complex market of interacting algorithms continues to evolve, it is also likely that existing algorithms that may have been innocuous yesterday will have significant impact tomorrow.

AI is subject to two significant types of bias:

- bias in its coding (both in design and development), or
- selection bias in or distortion/corruption of its data inputs.

Either type can result in significantly flawed results delivered under the patina of “independent” automated decision making.

The Criticality of Truly Applicable and Accurate Data Inputs

While much contemporary commentary has focused on the question of bias, the long experience of software development teaches that the proper scope, understanding and accuracy of data have dominant impacts on the efficacy of programming. In simple terms, “garbage in, garbage out.” This relationship is

[3] See discussion in Lohr (2017).

[4] For instance, media reports (see, for example, Wexler 2017) have pointed out clear racial bias resulting from reliance on sentencing algorithms used by many US courts.

particularly true with AI. AI is a process of machine learning — or, more accurately, machine teaching. The inaccuracies in data often come from reflections of human biases or human judgments about what data sets tell us. The establishment of training data and training features is at the heart of AI. As Rahul Barghava (2017) says, **"**In machine learning, the questions that matter are 'what is the textbook' and 'who is the teacher.' "The more scrutiny these can receive, the more likely that the data will be fit for purpose. To consider one example, some local governments in the United States have been making more use of algorithmic tools to guide responses to potential cases of children at risk. Some of the best implementations involve widespread academic and community scrutiny on their purpose, process and data. The evidence is that these systems can be more comprehensive and objective than the different biases people display when making high-stress screenings. But even then, the data accuracy problem emerges: "It is a conundrum. All of the data on which the algorithm is based is biased. Black children are, relatively speaking, over-surveilled in our systems, and white children are under-surveilled. Who we investigate is not a function of who abuses. It's a function of who gets reported."[5] Sometimes the data is just flawed. But the more scrutiny it receives,

the better it is understood. In the workplace, workers often have the customer and workflow experience to help identify such data accuracy challenges.

Acceptance of data inputs to AI in the workplace is not just a question of ensuring accuracy and fit for purpose. It is also one of transparency and proportionality.

The crisis surrounding Facebook, over Cambridge Analytica's illicit procurement of millions of its users' private data to inform data-targeting strategies in the 2016 US presidential election, has shown that there is a crisis in ethics and public acceptance in the data collection companies. Among the many issues raised by that scandal, a subset includes:

- a realization of the massive collection of data beyond the comprehension of the ordinary user;
- the corporate capacity to collate internal and external data and analyze it to achieve personally recognizable data profiles of users, which the users neither knew about nor explicitly approved;
- the collecting of people's data without any contractual or other authority to do so; and
- the lack of transparency in the data collection processes, sources, detail, purposes and use.

These issues are more urgent when they have a direct impact on people's working lives. It is important, to meet the pressing needs of data accuracy and worker confidence, that employees and contractors have access to the data being collected for enterprise AI, and, in particular, for workplace AI. Data quality improves when many eyes have it under scrutiny. Furthermore, to preserve their workplace morale, workers need to be sure that their own personal information is being treated with respect and in accordance with laws on privacy and labour rights.

Including Community Interests

The present discussion about the ethics of data gathering and algorithmic decision making has focused on the rights of individuals. The principles for the adoption of AI need to include an expression of the policy concerns of the community as a whole, as well as those of individuals. For instance, the individual right of intellectual property protection may need to be traded off against the community interest in non-discrimination (which is also an individual's human right) and, hence, a requirement for greater

[5] Erin Dalton, deputy director of Allegheny County's Department of Human Services, quoted in Hurley (2018).

transparency as to the purpose, as well as the inputs and outputs, of a particular algorithmic decision-making tool.

Risk of Further Marginalization of the Vulnerable

AI, at its heart, is a system of probability analysis for presenting predictions about certain possible outcomes. Whatever the use of different tools for probability analysis, the problem of outliers remains. In a world run by algorithms, the outlier problem has real human costs. A society-level analysis of the impact of big data and AI shows that their tendency toward profiling and limited-proof decisions results in the further marginalization of the poor, the Indigenous and the vulnerable (see Obar and McPhail 2018).

One account reported by Virgina Eubanks (2018, 11) explains how interrelated systems reinforce discrimination and can narrow life opportunities for the poor and the marginalized:

> What I found was stunning. Across the country, poor and working-class people are targeted by new tools of digital poverty management and face life-threatening consequences as a result. Automated eligibility systems discourage them from claiming public resources that they need to survive and thrive. Complex integrated databases collect their most personal information, with few safeguards for privacy or data security, while offering almost nothing in return. Predictive models and algorithms tag them as risky investments and problematic parents. Vast complexes of social service, law enforcement, and neighborhood surveillance make their every move visible and offer up their behavior for government, commercial, and public scrutiny.

This excerpt highlights the issue of unintended consequences, particularly costly when they impact the marginalized. It is unlikely that the code-writers of the systems described above started off with the goal "let's make life more difficult for the poor." However, by not appreciating the power of the outcome of the semi-random integration of systems — each system narrowly incented by the desired outcomes for the common and the privileged — that is exactly what these programmers did.

The same concerns apply to the workplace. As one example, at first glance it may appear intuitive to record how far an applicant lives from the workplace for an algorithm designed to determine more likely long-term employees. But this data inherently discriminates against poorer applicants dependent on cheaper housing and public transport. As another, AI written around a narrow definition of completed output per hour may end up discriminating against slower older employees, whose experience is not reflected in the software model.

Over the past few decades, many employers have adopted corporate social responsibilities, partly in the recognition that their contribution to society is more than just profitability. As the AI revolution continues, it is essential that a concerted effort be made to ensure that broader societal responsibilities are not unwittingly eroded through the invisible operation of narrowly written deterministic algorithms that reinforce each other inside and beyond the enterprise.

Big data and AI should not result in some sort of poorly understood, interlinked algorithmic Benthamism, where the minority is left with diminished life opportunities and further constrained autonomy.

Humans Are Accountable for AI

There is a tendency by some to view AI, because of its complex and opaque decision making, as being separate from other products made by humans, and a unified entity unto itself. Such a notion is a grave error and one that fails to understand the true role of the human within the algorithm. It is essential to emphasize the human agency within the building, populating and interpretation of the algorithm. Humans need to be held accountable for the product of algorithmic decision making. As Lorena Jaume-Palasí and Matthias Spielkamp (2017, 6-7) state:

> The results of algorithmic processes…are patterns identified by means of induction. They are nothing more than statements of probability. The patterns identified do not themselves constitute a conclusive judgment or an intention. All that patterns do is suggest a particular (human) interpretation and the decisions that follow on logically from that interpretation. It therefore seems inappropriate to speak of "machine agency", of machines as subjects capable of bearing "causal responsibility"...While it is true that preliminary automated decisions can be made by means of algorithmic processes (regarding the ranking of postings that appear on a person's Facebook timeline, for example), these decisions are the result of a combination of the intentions of the various actors who (co-)design the algorithmic processes involved: the designer of the personalization algorithm, the data scientist who trains the algorithm with specific data only and continues to co-design it as it develops further and, not least, the individual toward whom this personalization algorithm is directed and to whom it is adapted. All these actors have an influence on the algorithmic process. Attributing causal responsibility to an automated procedure — even in the case of more complex algorithms — is to fail to appreciate how significant the contextual entanglement is between an algorithm and those who co-shape it.

A Human-centric Model Is Essential for Acceptance of AI and to Ensure a Safe AI Future

Hundreds of technical and scientific leaders have warned of the risk of integrated networks of AI superseding human controls unless governments intervene to ensure human control is mandated in AI development. The British physicist Stephen Hawking spoke of the importance of regulating AI: "Unless we learn how to prepare for, and avoid, the potential risks, AI could be the worst event in the history of our civilization. It brings dangers, like powerful autonomous weapons, or new ways for the few to oppress the many" (quoted in Clifford 2017); further, he warned, "it would take off on its own, and re-design itself at an ever increasing rate. Humans, who are limited by slow biological evolution, couldn't compete, and would be superseded" (quoted by Cellan-Jones 2014).

More specifically within the workplace, big data and AI could result in a new caste system imposed on people by systems determining and limiting their opportunities or choices in the name of the code-writers' assumptions about the best outcome for the managerial purpose. One can imagine an AI-controlled recruitment environment where the freedom of the person to radically change careers is punished by algorithms only rewarding commonly accepted traits as being suitable for positions.

AI should not be allowed to diminish the ability of people to exercise autonomy in their working lives and in determining the projection of their own life paths. This autonomy is an essential part of what makes us human. As UNI Global Union (2018, 9) says, in the deployment of these technologies, workplaces should "show respect for human dignity [and] privacy and the protection of personal data should be safeguarded

in the processing of personal data for employment purposes, notably to allow for the free development of the employee's personality as well as for possibilities of individual and social relationships in the work place."

Microsoft (2018, 136) has called for a "human-centered approach" to AI. This approach is important not only to control AI's potential power, but to ensure — particularly in the workplace, including the gig economy — that AI serves the values and rights humans have developed as individuals in societies over the last centuries.

As *The Economist* (2018, 13) has concluded: "The march of AI into the workplace calls for trade-offs between privacy and performance. A fairer, more productive workforce is a prize worth having, but not if it shackles and dehumanises employees. Striking a balance will require thought, a willingness for both employers and employees to adapt, and a strong dose of humanity."

The Need to temper AI and related Big Data Manipulation of Users and Citizens

A long standing tenet of public policy in both advanced and emerging economies is that where an economic actor is in a position to manipulate a consumer – in a position to exploit the relative vulnerabilities or weaknesses of a person in order to usurp their decision making– society requires their interests to be aligned and punishes acts that are seen as out of alignment of the interests of the person. Individuals in some relationships, for example between priests-parishioners, lawyers-clients, doctors-patients, teachers-students, therapists-patients, etc., are vulnerable to manipulation through the intimate data collected by the dominant actor, and these types of relationships are governed such that the potential manipulator is expected to act in accordance with the interests of the vulnerable party. We regularly govern manipulation that undermines choice, such as when negotiating contracts under duress or undue influence, or when contractors act in bad faith, opportunistically, or unconscionably. The laws in most countries void such contracts.

When manipulation works, the target's decision making is usurped to pursue the interests of the manipulator; and the tactic is never known by the target. Some commentators rightly compare manipulation to coercion (Susser, Roessler, and Nissenbaum 2019). For coercion, a target's interests are overtly overridden by force and the target knows about the threat and coercion. Manipulation, on the other hand, overrides a target's choice subversively. Both seek to overtake the authentic choice of the target and just choose different tactics. In this way, manipulation has the goals of coercion and the deception of fraud. And offline, we regulate manipulation similar to the way we regulate coercion and fraud: to protect consumer choice-as-consent and preserve the autonomy of the individual.

Online actors, such as data aggregators, data brokers, and ad networks, can not only predict what we want and how badly we need it but can also leverage knowledge about when an individual is vulnerable to making decisions in the interest of the firm. Recent advances in hyper-targeted marketing allows firms to generate leads, tailor search results, place content, and develop advertising based on a detailed picture of their target. Aggregated data on individuals' concerns, dreams, contacts, locations, and behaviors allows marketers to predict what consumers should want and how to best sell to them. It allows firms to predict moods, personality, stress levels, health issues, etc. – and potentially use that information to undermine

the decisions of consumers. In fact, Facebook recently offered advertisers the ability to targets teens when they are 'psychologically vulnerable.'

All this information asymmetry between users and data aggregators has sky-rocketed in recent years.

The data collection industry is not new. Data brokers like Acxiom and ChoicePoint have been aggregating consumers' addresses, phone numbers, buying habits and more from offline sources and selling them to advertisers and political parties for decades. But the Internet has transformed the space. The scope and intimacy of the data collection and the purposes for which it is sold and used is rarely comprehended by users.

One reason for this is that much of the data is collected in a non-transparent way and mostly in a manner that people would not consider covered by contractual relationships. Many Internet users, at least in developed countries, have some understanding that the search engines and the e-commerce engines collect data on what sites they have visited and that this data is used to help tailor advertising to them. But most have little idea of just how extensive this commercial surveillance is. A recent analysis of the terms and conditions of the big US platforms shows that they collect 490 different types of data on each user.[6]

A recent study of 1 million web sites showed that nearly all of them allow third party web trackers and cookies to collect user information to track page usage, purchase amounts, browsing habits, etc. Trackers send personally identifiable information such as user's name, address, and email and spending details. These latter allow the data aggregators to then de-anonymize much of the data they collect (Englehardt and Narayanan 2016, Libert, 2015).

But cookies are only one of the mechanism used to collect data on people. Both little known data aggregators and the big platforms draw huge amounts of information from cell towers, the use of the devices themselves, many of the third party apps running on the user's device, Wi-Fi access, as well as public data sources and third party data brokers.

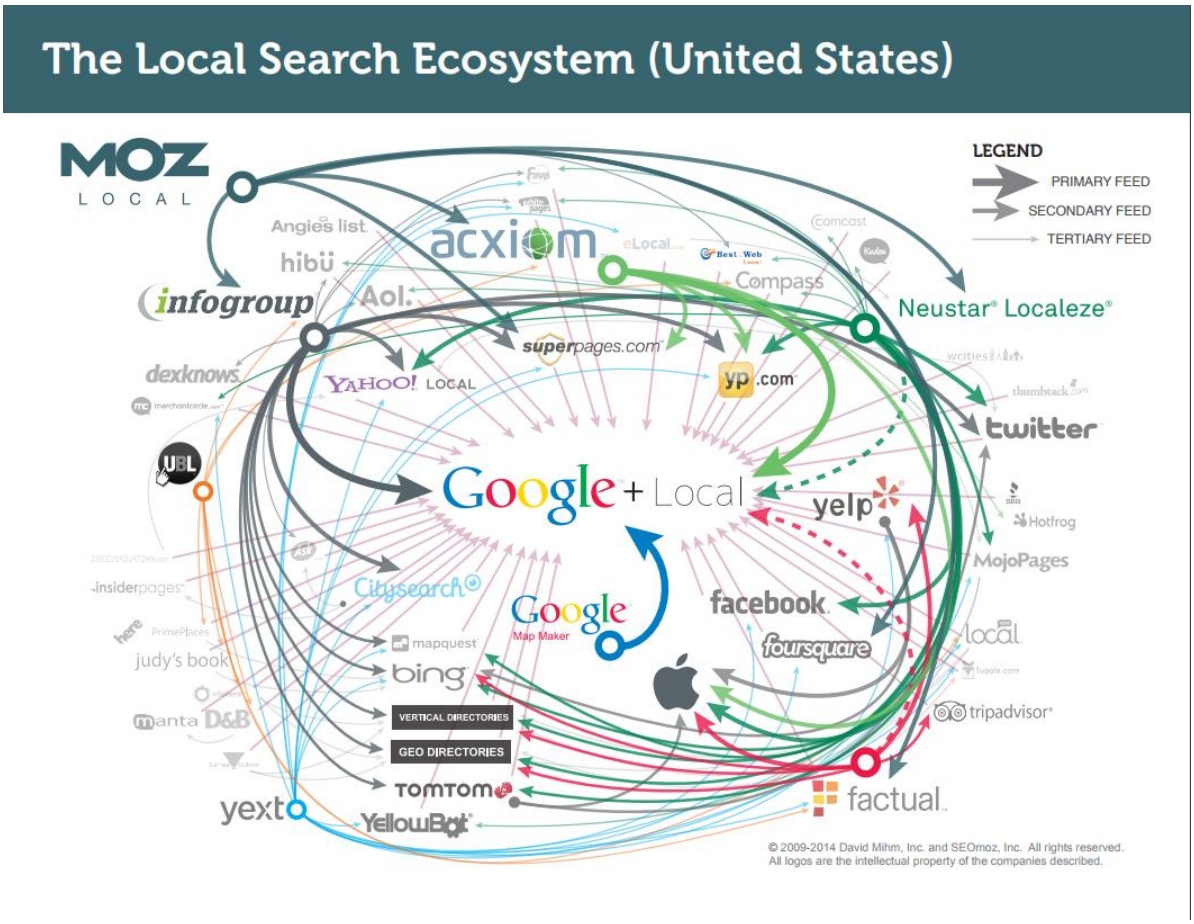
As the New York Times recently reported:

> Every minute of every day, everywhere on the planet, dozens of companies — largely unregulated, little scrutinized — are logging the movements of tens of millions of people with mobile phones and storing the information in gigantic data files. The Times Privacy Project obtained one such file [which] holds more than 50 billion location pings from the phones of more than 12 million Americans as they moved through several major cities... Each piece of information in this file represents the precise location of a single smartphone over a period of several months…It originated from a location data company, one of dozens quietly collecting precise movements using software slipped onto mobile phone apps.[7]

[6] See the publicly available data at https://mappingdataflows.com/

[7] "One nation, tracked An investigation into the smartphone tracking industry from Times Opinion" https://www.nytimes.com/interactive/2019/12/19/opinion/location-tracking-cell-phone.html?searchResultPosition=8

An indication of the scale and complexity of the collection and transfer of user data among web sites can be gleaned from the following diagram. Devised by David Mihm, a noted expert on search engine optimization, it shows the data feeds contributing to the US online local search ecosystem.[8]



It is data collection networks and markets like these, invisible to the vast majority of the people whose personal data is being collected, which enable Cambridge Analytica (of the 2016 US Presidential election fame) to claim that it holds to have up to five thousand data points on every adult in the US.[9]

AI in the military

[8] https://whitespark.ca/blog/understanding-2017-u-s-local-search-ecosystem/

[9] See "MPs grill data boss on election influence", 27 February 2018 http://www.bbc.com/news/technology-43211896

In 2015, a group of leading AI researchers and investors signed an open letter warning of the dangers of autonomous weapons. "The key question for humanity today is whether to start a global AI arms race or to prevent it from starting. If any major military power pushes ahead with AI weapon development, a global arms race is virtually inevitable."[10] Today, many nations are pushing to apply AI for military advantage. While the phrase "AI arms race" is misleading – AI is a general technology enabler rather than a weapons system in itself – the rush to deploy it brings with it real risks. As Paul Scharre has written, "The widespread adoption of military AI could cause warfare to evolve in a manner that leads to less human control and to warfare becoming faster, more violent, and more challenging in terms of being able to manage escalation and bring a war to an end. Additionally, perceptions of a "race" to field AI systems before competitors do could cause nations to cut corners on testing, leading to the deployment of unsafe AI systems that are at risk of accidents that could cause unintended escalation or destruction."[11]

Partly in response to at least some of these concerns, the US Department of Defense adopted in 2020 a set of AI ethical principles encompassing five major areas:

1. Responsible. DoD personnel will exercise appropriate levels of judgment and care, while remaining responsible for the development, deployment, and use of AI capabilities.
2. Equitable. The Department will take deliberate steps to minimize unintended bias in AI capabilities.
3. Traceable. The Department's AI capabilities will be developed and deployed such that relevant personnel possess an appropriate understanding of the technology, development processes, and operational methods applicable to AI capabilities, including with transparent and auditable methodologies, data sources, and design procedure and documentation.
4. Reliable. The Department's AI capabilities will have explicit, well-defined uses, and the safety, security, and effectiveness of such capabilities will be subject to testing and assurance within those defined uses across their entire life-cycles.
5. Governable. The Department will design and engineer AI capabilities to fulfill their intended functions while possessing the ability to detect and avoid unintended consequences, and the ability to disengage or deactivate deployed systems that demonstrate unintended behavior.[12]

But when the issue of limiting the development of autonomous weapons systems has arisen for international discussion, there has been no consensus on legal action to limit their use. Such a limitation could be achieved through a new protocol to the Convention on Conventional Weapons (CCW), which has discussing this concern since 2014. While most states have recognised the need to retain some form of human control over these weapons, neither the United States nor Russia is willing to enter yet into negotiations of a limitations agreement.

---

[10] "Autonomous Weapons: An Open Letter from AI & Robotics Researchers," Future of Life Institute, 2015, https://futureoflife.org/open-letter-autonomous-weapons/?cn-reloaded=1.

[11] Paul Scharre, "Debunking the AI race myth", *Texas National Security Review*, Volume 4, Issue 3, (Summer 2021) pp 121-132, at p 122.

[12] https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/

2 Impact on Rights

Protection against discrimination[13]

Our colleague Dr. Jutta Treviranus has written that one area that requires further emphasis is the tendency for machine learning to amplify, accelerate and automate existing discrimination against minorities and outliers. This inevitably occurs even when there is no bad intent or bias on the part of the developers or implementers. It is a diffuse effect that exponentially increases with each iteration and with each machine learning training session. This is not addressed by ensuring proportional representation and closing data gaps. It is not addressed by current AI ethics auditing tools that compare treatment of a specific protected identity group with treatment of the population as a whole,[14] or elimination of obvious human bias. The problem pre-dates AI, and big data analytics. The origin is in the principle of majority rules, evidence based on statistical reasoning and decisions based on probability. Prior to big data and AI there was a greater possibility of the determination of reasonable exceptions. Automated decisions remove this possibility and current AI ethics measures only make it harder to argue mistreatment. The impact is exponentially accelerated and amplified.

One group that feels the impact most is people with disabilities.[15] The only common data footprint of disability is sufficient difference from the mean that systems do not address your needs. Disability is at the outer edge of every justice seeking group, at the same time it is the world's largest minority.[i16]It has no bounded definition and because of stigma associated with disability, people often do not self-identify. Many people with disabilities represent an 'n' of one. People with disabilities tend to be the extreme small minorities and outliers in any population data set. The impacts of disability complexly compound through all aspects of life, including poverty, education, work, health, digital inclusion, etc. Scientific evidence and determinants of scientific "truth" for the general population tends to be wrong if you have a disability. Decisions guided by data will likely rule against you. The extreme injustice of this is that the more critical your needs, and the better off the general population is, the more likely your needs will not be addressed. The trivial needs of the many will overpower the critical needs of the few in majority-rules decision systems. The accelerating drift of data sets means the situation worsens as AI becomes more pervasive.

People with disabilities are most vulnerable to data abuse and misuse and current privacy protections do not work. Because of the uniqueness of the data footprint, people with disabilities can be easily reidentified with any aggregation of data. At the same time differential privacy removes the unique data needed to serve the needs of someone with a disability.

---

[13] My thanks to Jutta Treviranus for this section on discrimination.

[14] https://www.brookings.edu/research/auditing-employment-algorithms-for-discrimination/

[15] Treviranus, J., Gupta, A., (2020). Inclusive Designed Artificial Intelligence. In Schaffers, H. Vartianen, M., Bus, J., *Digital Innovation and Societal Change*. River Publishers, London, UK.

[16] Trewin, S., Basson, S., Muller, M., Branham, S., Treviranus, J., Gruen, D., Hebert, D., Lyckowski, N. and Manser, E., 2019. Considerations for AI fairness for people with disabilities. AI Matters, 5(3), pp.40-63.

There is a discriminatory hierarchy even within disability. The more you are at the margins, the harder it is to use personalization systems intended to meet your needs. Every learning model tends toward a mean. To train the model to serve your unique needs requires a great deal of "swimming upstream." Thus "struggling students" are less likely to be served by instructional tutors.[17] Individuals whose speech differs greatly from the norm have greater difficulty using speech recognition systems and people who are blind but live in greater poverty will not be able to use pattern recognition systems intended to replace vision.

(The beneficial examples listed in the paper do not address the bias against disability in hiring systems. Behavioural science has many of the built-in biases of other data analytics systems. Even if the assessments are made accessible in systems such as "Pymetrics," the accessible systems will be more complex and involve more cognitive load.)

The only means to address the discriminatory drift is to support bottom-up data with user-controlled post-hoc aggregation.[18] To protect privacy, intelligence would be on the personal device. Individuals and communities would co-create the data applications. Data aggregation could be stewarded by cooperative data trusts.[19]

A compromise measure would be mandatory periodic data refresh, to fight both dataset pollution and the discriminatory drift. The application of AI to assist in decisions or filter individuals should require a disability impact assessment. A disability impact assessment will be an indicator of treatment of all forms of difference. Data nutrition labels[20] and systems that signal when AI guidance is likely not to apply within a given decision instance would help to reduce the risk of false positive security decisions or critical decisions in health or education.

The current shortcomings of even ethical AI, namely difficulty addressing the unexpected, and difficulty transferring to new contexts, would benefit from more diversity-supportive decision systems. The greatest diversity is at the margins of a data set. Thus, addressing the barriers and discrimination against people with disabilities would be beneficial to the entire population.

Competition Issues

Barriers to entry exist in many AI driven digital markets due to network effects (the value of services rise with the number of users), first mover and scale advantages in accumulating machine learning training data sets, and a range of other factors. The treatment of the resulting power asymmetries should be treated

[17] Treviranus, J. (2021). Learning to Learn Differently. In: Holmes, W. (ed) *Ethics of AIED. Who Cares?*" Taylor & Francis Group, Oxon, UK.
[18] Tchernavskij, P. (2019). Designing and Programming Malleable Software (Doctoral dissertation, Université Paris-Saclay (ComUE)).
[19] https://ec.europa.eu/jrc/communities/sites/jrccties/files/micheli.pdf
[20] https://datanutrition.org/

more analogously to the regulation of natural monopolies offline.[21] Many jurisdictions, including the EU (via the Digital Services Act), have begun the process of investigations and accompanying legislative reform to re-establish the conditions for effective competition in these markets.

Many have also promoted the collection of large data sets under public commons rules to achieve a number of objectives:

- restore agency to people regarding how their data is circulated and used – recognising that people often have a range of desires that include but are not limited to the commercial sphere,
- help to increase the amount and quality of data potentially available to all firms, not just the largest technology platforms, to build a more level playing field, boosting competition in line with competitiveness goals, values and fundamental rights.

One country where authorities are moving to reduce big platform/AI players' market power is China. The ongoing shake-up of the Tech sector is partly driven by a sense that big data sets should be more freely available to smaller players in the market, as well as competing state owned enterprises. As *The Economist* noted on 14 August 2021:

> As a guiding principle, the vice-premier, Liu He, recently stated that China is moving into a new phase of development that prioritises social fairness and national security, not the growth-at-all-costs mentality of the past 30 years…
>
> Start with data. Europe and some American states, such as California, have devised laws that seek to protect consumers from the misuse of their personal information by large companies. China has put similar rules in place; in some cases they are more severe than in the West. But Chinese regulators are going further. In a largely ignored, jargon-filled policy paper from the State Council, China's cabinet, in April last year, data were named as a "factor of production" alongside capital, labour, land and technology…
>
> China's new data policy remains a work in progress. The Data Security Law will come into force on September 1st and the Personal Information Protection Law is due to be adopted by China's rubber-stamp parliament soon. It is unclear how they will be enforced, though data specialists intuit that many types of data currently held by internet giants could eventually be traded on government-backed and private exchanges. Ant, for example, is already being prodded by authorities to open up its vast stores of personal financial data to state-owned companies and smaller tech rivals.[22]

---

21 For a summary of the literature on the regulation of natural monopolies, see Joskow (2007). For a recent analysis, see Ducci (2020).

22 "What Tech does China want", *The Economist*, August 14, 2021. https://www.economist.com/business/what-tech-does-china-want/21803410

AI and Human Rights

Much of the public critique of AI focuses on racial or sex bias in learning data sets and or the outputs of the systems. But the right to non-discrimination is not the only human right vulnerable to AI deployments. Indeed broadly deployed AI affects nearly every internationally recognized human right, from the rights to privacy and freedom of expression, to the rights to health and education.[23]

Accessnow, an international NGO focused on the digital rights of users at risk around the world, has issued a review of how human rights are challenged by AI and makes the following broad recommendations:

1. Comprehensive data protection legislation can anticipate and mitigate many of the human rights risks posed by AI. However, because it is specific to data, additional measures are also necessary.
2. Government use of AI should be governed by a high standard, including open procurement standards, human rights impact assessments, full transparency, and explainability and accountability processes.
3. Given the private sector's duty to respect and uphold human rights, companies should go beyond establishing internal ethics policies and develop transparency, explainability, and accountability processes.
4. Significantly more research should be conducted into the potential human rights harms of AI systems and investment should be made in creating structures to respond to these risks.[24]

Much detailed discussion of the effect of AI on human rights (at least as they have evolved in a European context) is contained in the work of the Ad Hoc Committee on Artificial Intelligence at the Council of Europe[25] (CAHAI), which has in its Feasibility Study[26] addressed both the need for international regulation, as well as the instruments through which this could be achieved - identifying a need for a combination of binding and non-binding instruments, some with horizontal character, some vertical or sectorial.

Some Framework Principles

Over the last several years, the author has developed (in response to the G20 policy processes) a series of principles which give a framework for how governments could seek to protect the rights and well-being of citizens and workers in AI infused world. Partly to make the principles more palatable across a range of political systems, the author has cast many of the principles mostly in the context of the future of work (rather than broader civic life). Every government, not just liberal democracies, is confronted with the challenge of AI in the workplace and the need to maintain worker confidence.

---

[23] The rights being referred to are protected in the three main treaties: the Universal Declaration of Human Rights (UDHR), the International Covenant on Civil and Political Rights (ICCPR), and the International Covenant on Economic, Social and Cultural Rights (ICESCR).

[24] See *Human Rights In The Age Of Artificial Intelligence* https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights

[25] See https://www.coe.int/en/web/artificial-intelligence/home

[26] See https://rm.coe.int/cahai-2020-23-final-eng-feasibility-study-/1680a0c6da

The first set of seven framework principles relates to the collection of data in the work environment.

Right to know data is being collected, for what and from where: Workers, be they employees or contractors, or prospective employees and contractors, must have the right to know what data is being collected on them by their employers, for what purpose and from what sources.

Right to ensure worker data is accurate and compliant with legal rights to privacy: An important feature for worker understanding and productivity is to ensure that workers, ex-workers and job applicants have access to the data held on them in the workplace or have the means to ensure that the data is accurate and can be rectified, blocked or erased if it is inaccurate or breaches legally established rights to privacy. The collection and processing of biometric data and other personally identifiable information (PII) must be proportional to its stated purpose, based on scientifically recognized methods, and held and transmitted very securely.

Principle of proportionality: The data collected on present or prospective employees or contractors should be proportional to its purpose. As one group has proposed: "Collect data and only the right data for the right purposes and only the right purposes, to be used by the right people and only the right people and for the appropriate amount of time and only the appropriate amount of time."

Principle of anonymization: Data should be anonymized where possible. Data with PII should only be available where it is important to the data collection's prime purpose, and its visibility must be limited to the employee and the relevant manager. Aggregated, anonymized data is preferable for many management and productivity purposes.

Right to be informed about the use of data: Employees and contractors should be fully informed when either internal or external data (or both) has been used in a decision affecting their career. Any data processing of present or prospective employees' or contractors' data should be transparent and the available for their review. The right to understand and appeal against both the rationale employed and the data used to achieve that rationale is essential to safeguard present or prospective workers against poor or inaccurate input data or discriminative decisions.

Limits to monitoring of the workplace by employers: Proportional data collection and processing should not be allowed to develop into broad-scale monitoring of employees or contractors. While monitoring can be an indirect consequence of steps taken to protect production, health and safety or to ensure the efficient running of an organization, continuous general monitoring of workers should not be the primary intent of the deployment of workplace technology. Given the potential in the use of such technology to violate the rights and freedoms of the persons concerned, employers must be actively engaged to ensure that the use is constrained to specific positive purposes, so as not to breach these rights. This principle is not only a matter of workplace freedoms, but also a practical step toward maintaining morale and productivity.

Accuracy of data inputs and the "many eyes" principle: Employers should ensure the accuracy, both in detail and its intended purpose, of the data models and sources for AI. Poor data results in flawed decision

making. Training data and training features should be reviewed by many eyes to identify possible flaws and to counter the "garbage in, garbage out" trap. There should be a clear and testable explanation of the type and purpose of the data being sourced. Workers and contractors with experience of the work processes and data environment of the firm should be incorporated into the review of data sources. Such data should be regularly reviewed for accuracy and fit for purpose. Algorithms used by firms to hire, fire and promote should be regularly reviewed for data integrity, bias and unintended consequences.

An additional seven principles focus on AI in the workplace.
Focus on humans: focus: Human control of AI should be mandatory and testable by regulators.
AI should be developed with a focus on the human consequences as well as the economic benefits. A human impact review should be part of the AI development process, and a workplace plan for managing disruption and transitions should be part of the deployment process. Ongoing training in the workplace should be reinforced to help workers adapt. Governments should plan for transition support as jobs disappear or are significantly changed.

Shared benefits: AI should benefit as many people as possible. Access to AI technologies should be open to all countries. The wealth created by AI should benefit workers and society as a whole as well as the innovators.

Fairness and inclusion: AI systems should make the same recommendations for everyone with similar characteristics or qualifications. Employers should be required to test AI in the workplace on a regular basis to ensure that the system is built for purpose and is not harmfully influenced by bias of any kind — gender, race, sexual orientation, age, religion, income, family status and so on. AI should adopt inclusive design efforts to anticipate any potential deployment issues that could unintentionally exclude people. Workplace AI should be tested to ensure that it does not discriminate against vulnerable individuals or communities.
Governments should review the impact of workplace, governmental and social AI on the opportunities and rights of poor people, Indigenous peoples and vulnerable members of society. In particular, the impact of overlapping AI systems toward profiling and marginalization should be identified and countered.

Reliability: AI should be designed within explicit operational requirements and undergo exhaustive testing to ensure that it responds safely to unanticipated situations and does not evolve in unexpected ways. Human control is essential. People-inclusive processes should be followed when workplaces are considering how and when AI systems are deployed.

Privacy and security: Big data collection and AI must comply with laws that regulate privacy and data collection, use and storage. AI data and algorithms must be protected against theft, and employers or AI providers need to inform employees, customers and partners of any breach of information, in particular PII, as soon as possible.

Transparency: As AI increasingly changes the nature of work, workers, customers and vendors need to have information about how AI systems operate so that they can understand how decisions are made. Their involvement will help to identify potential bias, errors and unintended outcomes. Transparency is not necessarily nor only a question of open-source code. While in some circumstances open-source code will

be helpful, what is more important are clear, complete and testable explanations of what the system is doing and why.
Intellectual property, and sometimes even cyber security, is rewarded by a lack of transparency. Innovation generally, including in algorithms, is a value that should be encouraged. How, then, are these competing values to be balanced?
One possibility is to require algorithmic verifiability rather than full algorithmic disclosure. Algorithmic verifiability would require companies to disclose not the actual code driving the algorithm but information allowing the *effect* of their algorithms to be independently assessed. In the absence of transparency regarding their algorithms' purpose and actual effect, it is impossible to ensure that competition, labour, workplace safety, privacy and liability laws are being upheld.[27]

When accidents occur, the AI and related data will need to be transparent and accountable to an accident investigator, so that the process that led to the accident can be understood.

A related principle is data governance of record keeping: Long term data governance throughout the AI system lifecycle should be required to ensure that data used in AI systems is accurate, complete and appropriate and is stored in a safe and secured environment. Further appropriate records of the data management methodologies should be maintained.

Accountability: People and corporations who design and deploy AI systems must be accountable for how their systems are designed and operated. The development of AI must be responsible, safe and useful. AI must maintain the legal status of tools, and legal persons need to retain control over, and responsibility for, these tools at all times.

Workers, job applicants and ex-workers must also have the "right of explanation" when AI systems are used in human-resource procedures, such as recruitment, promotion or dismissal.[28] They should also be able to appeal decisions by AI and have them reviewed by a human.

Sustainability. AI should be able to detect unintended environmental harm and automatically disengage if it occurs, or allow deactivation by a human. It is particularly important that AI and autonomous devices deployed in agriculture and mining should be designed and monitored for long term environmental sustainability and maintenance of biodiversity. Leaving agriculture just in the thrall of the efficiency motive will result in monocultures and loss of food diversity.

Principles to protect the citizen as consumer as well as worker

In the offline world, we have developed safeguards to ensure that those with intimate knowledge of others do not exploit vulnerabilities and weaknesses of individuals through manipulation. Yet, online data

[27] This is explored to some degree by the Global Commission for Internet Governance (2016, 45).

[28] The European Union's General Data Protection Regulation seems to infer a "right to explanation." See Burt (2017).

aggregators and their related AI firms, with whom we have no relationship (for instance a contract), have more information about our preferences, concerns, and vulnerabilities than our priests, doctors, lawyers, or therapists. I propose that Governments should extend their existing off-line protections and standards against manipulation to also cover these data controllers which presently have the knowledge and proximity of a very intimate relationship without the governance and trust inherent to such relationships in the off-line market. I also propose several steps to protect citizens' autonomy and decrease user deception.

Regulating manipulation to protect consumer choice is not novel. What is unique now is that the current incarnation of manipulation online divorces the intimate knowledge of the target and power used to manipulate from a specific, ethically-regulated relationship as we usually find offline. Online we now have a situation where firms, with whom we have no relationship, have more information about our preferences, concerns, and vulnerabilities than our priests, doctors, lawyers, or therapists. In addition, these firms, such as ad networks, data brokers, and data aggregators, have an ability to reach specific targets due to the hypertargeting mechanisms available online. Yet, we are not privy to who has access to that information when businesses approach us with targeted product suggestions or advertising. These data brokers have the knowledge and proximity of an intimate relationship covering very personal parts of our lives without the governance and trust inherent to such relationships in the market. They clearly fail the transparency, stewardship, non-discrimination, autonomy, and fairness provisions of the G20 Principles.

Current Approach to Regulating Manipulation Online.

In the offline world sharing information with a particular market actor, such as a firm or individual, requires trust and other safeguards such as regulation, professional duties, contracts, negotiated alliances, nondisclosure agreements, etc. The point of such instruments is to share information within a (now legally binding) safe environment where the interests of the two actors are forced to be aligned. However, three facets of manipulation by data traffickers[29] – those in a position to covertly exploit the relative vulnerabilities or weaknesses of a person in order to usurp their decision making – strain our current mechanisms governing privacy and data. First, manipulation works by not being disclosed, thus making detection difficult and rendering the market ill-equipped to govern the behavior. Second, the type of manipulation described herein is performed by multiple economic actors including websites/apps, trackers, data aggregators, ad networks, and customer facing websites luring in the target. Third, data traffickers – who collect, aggregate, and sell consumer data – are the engine of manipulation of online consumers yet have no interaction, contract, agreement with individuals.

These three facets – manipulation is deceptive, shared between actors, and not visible by individuals – render the current mechanisms ineffective in governing the behavior or the actors. For example, GDPR is strained when attempting to limit a 'legitimate use' of data traffickers or data brokers who are looking to market products and services based on intimate knowledge. An individual has a right to the restriction of processing of information only when there are no legitimate grounds of the data controller. This makes GDPR fall short because legitimate interests can be broadly construed to include product placements and

[29] Lauren Scholz first used the term data traffickers, rather than data brokers, to describe firms that remain hidden yet traffic in such consumer data (Scholz 2019).

ads. And the manipulation of individuals has not been identified (yet) as diminishing a human right of freedom and autonomy. One fix is to more clearly link manipulation to individual autonomy, which would be seen as a human right that could trump even the legitimate interests of data traffickers.

A first step forward – Policy Goals

In general, the danger comes from using intimate knowledge about an individual and hyper-targeting to then manipulate them. The combination of individualized data and individualized targeting needs to be governed or limited:

1. **Protect Autonomy.** Manipulation is only possible because a market actor, here it is data brokers, has intimate knowledge of individuals as to what renders a target vulnerable in their decision making. The goal of governance would be to limit the use of intimate knowledge by making certain types of intimate knowledge either illegal or heavily governed. The combination of intimate knowledge with hyper- targeting of individuals should be more closely regulated than blanket targeting based on age and gender. Explicitly recognize individual autonomy, defined as the ability of individuals to be the authentic authors of their own decisions, as a legal right in order to protect individuals from manipulation done in the name of "legitimate interests" within the AI Principles.
2. **Expand Definitions of Intimate Knowledge.** One step would be to explicitly include inferences made about individuals as sensitive information within such existing regulations as GDPR (Wachter and Mittelstadt 2019). Sandra Wachter and Brent Mittelstadt have recently called on rights of access, notification, and correction for not only the data being collected but the possible inferences drawn from the data about individuals. These inferences would be considered intimate knowledge of individuals that could be used to manipulate them (e.g., whether someone is depressed or not based on their online activity). The inferences made by data traffickers based on a mosaic of information about individuals can constitute intimate knowledge as to who is vulnerable and when. Current regulatory approaches only include collected data as protected rather than the inferences drawn about individuals based on that data.
3. **Force Shared Responsibility.** Make customer-facing firms responsible for who they partner with to track users or to target users. Customer-facing websites and apps should be responsible for who is given access to their users' data – whether by sale or whether given access by placing trackers and beacons on their site. Third parties include all trackers, beacons, and third parties who purchase data or access to their users. Websites and apps would then be held responsible for whether they partner with firms that abide by GDPR standards, AI Principles, or new standards of care in the U.S. Holding customer facing firms responsible for how their partners (third party trackers) gather and use their users' data would be similar to holding a hospital responsible for how the patient is cared for by their contractors in the hospital or holding a car company responsible for a third party app in the car that then tracked your movements. This would force the customer-facing firm, with whom the individual has some influence, to make sure their users' interests are being respected.[30] The shift would be to have customer-facing firms be held responsible for how their partners (ad networks and media) treat their users.

[30] It is ironic that currently data traffickers can *sell* data to bad actors but they just can't have their data *stolen* by those same bad actors.

4. Expand the Definition of "Sold". Make sure all regulations include beacons and tracking companies in the any requirement to notify if user data is 'sold'.
5. Create a Fiduciary Duty for Data Brokers. there is a profound, yet relatively easy to implement, step to address this manipulation. G20 and other governments could make their AI Principles practical by extending the regulatory requirements they have on doctors, teachers, lawyers, government agencies and others who collect and act on the intimate data of individuals to also apply to data aggregators and their related AI implementations. Any actor who collects intimate data about an individual should be required to act on, share, or sell this data consistent with the interests of the person. This would force the alignment of interests between the target/consumer/user and the firm in the position to manipulate. Without any market pressures, data brokers who hold intimate knowledge of individuals, would need to be held to a fiduciary-like standard of care for how their data would be used.(Balkin 2015) This would mean data brokers would need to be responsible for how their products and services were used to possibly undermine the interests of the individuals.
6. Add Oversight. Add a GAAP-like governance structure over data traffickers and ad networks to ensure individualized data is not used to manipulate. With these economic actors well outside any market pressures, there are few pressures on the firms to align their actions with users' interests. A third step would be to make data traffickers abide by GAAP-like regulations. Recently McGeveran called for GAAP-like approach for data security, where companies would be held to a standard defined for all firm similar to the use of GAAP standards for accounting. However, the same concept should be applied to those who hold user data as to how they protect the data when profiting from it.[31] Audits could also be used in order to ensure data traffickers, who control and profit from intimate knowledge of individuals, are abiding by their standards. This would add a cost to those who traffic in customer vulnerabilities and provide a third party to verify that those holding intimate user data act in a way that is in the individuals' interests and protect firms from capitalizing on their vulnerabilities. A GAAP-line governance structure could be flexible enough to understand the market needs while still being responsive to protect individual rights and concerns.
7. Decrease Deception. Finally, manipulation works because the tactic is hidden from the target. The goal of governance would be to make the basis of manipulation open to the target and others. In other words, make the type of intimate knowledge used in targeting obvious and public. This could mean a notice (e.g., this ad was placed because the ad network believes you are diabetic) or this could mean a registry when hyper-targeting is used to allow others to analyze how and why individuals are being targeted. Registering would be particularly important for political advertising so that researchers and regulators can identify the basis for hyper-targeting. It should not be sufficient for an AI/data aggregator just to say "I am collecting all this information in the interests of the user to see tailored advertising." That is equivalent to a doctor saying "I collect all this data about a patient's health to ensure that the patient only knows the prescription I give the patient." Patients have to give permission for and are entitled to know what data is collected (indeed in many countries patients formally own their health data), what tests have been conducted and their results, what the diagnosis is – and they are entitled to a second opinion on the data. Similar sorts of transparency and accountability offline should apply online. In other areas, where a lawyer or

[31] McGeveran calls for a GAAP like approach for data security. Here we would have the same idea for data protection. Where standards are set and others must be certified to abide by them (McGeveran 2018).

realtor or financial advisor, has intimate knowledge and a conflict of interest (where they could profit in a way that is detrimental to their client), they must disclose their conflict and the basis for their conflict.

Putting a legal requirement for companies to use data in the interests of the data subject also demands an objective test to ensure that the interpretation of the "interests of the data subject" is not open to differing interpretations. Various entities and companies could claim to be acting in the individual's interest, as they define it, even if the individual believes they are not. We propose that the test be grounded in two existing bodies of law: conventions on human rights and law governing relationships between professionals and their data subjects (doctor-patient, lawyer-client etc.), particularly the law related to use of patient/client data so as not to manipulate or exploit the data subject.[32]

The same principle holds for data that is generated by material objects owned by the data subject. The IOT digital service provider, when different from the owner of the material objects, are to manage the IOT data flow in the interests of the data subject and the data subject needs to be given automatic access to the data generated by the relevant material objects. This data, along with associated terms and conditions, must be transparent and clear.

In the offline world, we have stressed the importance of clear relationships between people and those who have intimate information asymmetries over them. And we have developed safeguards to ensure that those gaining positions of power do not exploit vulnerabilities and weaknesses of individuals. The issues posed by vast data collection and hyper-targeted marketing and/or service delivery are a product of the global expanse of the Internet, social media and AI platforms. Furthermore, the ability of 'data traffickers' and their AI partners to leverage knowledge they have on almost every person on the Internet makes the scale of the public policy and political challenge worthy of Ministers and Heads of Government. As the growing "tech backlash" shows, there is political mobilization among citizens across the world for change. The innovation of this "apply the offline world rules to the online players" approach is that it does not require governments to educate or force citizens to change behaviors or desires. It puts the ethical and regulatory onus on the firms involved and holds them accountable.

Multilateral Governmental Responses to Date

The questions of the correct governance for Artificial Intelligence and its underlying Big Data have been discussed at national and dispersed international fora for several years. These include efforts by the Council

32 Some examination of this law can be found at https://ec.europa.eu/health/sites/health/files/cross_border_care/docs/2018_mapping_patientsrights_frep_en.pdf

of Europe[33], the Innovation Ministers of the G7[34] , the European Parliament [35] and the OECD.[36] In June 2019, China's Ministry of Science and Technology published on its website the Governance Principles for a New Generation of Artificial Intelligence: Develop Responsible Artificial Intelligence.[37] The same month, the G20 Trade Ministers and Digital Economy Ministers adopted a set of AI Principles[38] which drew from the OECD's principals and discussion of proposals from G20 engagement groups[39] These principles point to a more human-focused and ethical approach to guiding AI – but they are by necessity broad in tone and lacking in regulatory specifics. The G20 principles are attached as an Appendix A.

On 11 June 2020, United Nations Secretary General Guterres presented his *Roadmap on Digital Cooperation*.[40] One of the recommended actions is "Supporting global cooperation on artificial intelligence that is trustworthy, human-rights based, safe and sustainable and promotes peace."

In April 2021, the European Commission released its Proposal for consideration of the European Parliament and Council for the promotion and regulation of AI in Europe.[41] The media brief outlining the full package is attached as Appendix B

The proposal is more detailed than previous statements of principles. Among a range of issues, the draft regulations seek to cover facial recognition, autonomous driving, the use of AI in online advertising, automated hiring, and credit scoring. They seek to prohibit (at least in some ways) "high risk" applications of AI, including law enforcement real time use of AI for facial recognition in public spaces (but not its post-facto uses in a number of circumstances).

---

[33] https://rm.coe.int/algorithms-and-human-rights-study-on-the-human-rights-dimension-of- aut/1680796d10

[34] https://g7.gc.ca/en/g7-presidency/themes/preparing-jobs-future/g7-ministerial-meeting/chairs- summary/annex-b/

[35] Directorate-General for Parliamentary Research Services (European Parliament), A governance framework for algorithmic accountability and transparency see at https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU(2019)624262_EN.pdf

[36] https://www.oecd.org/going-digital/ai/principles/

[37] See https://perma.cc/V9FL-H6J7

[38] See Annex to G20 Ministerial Statement on Trade and Digital Economy at https://www.mofa.go.jp/files/000486596.pdf

[39] For instance, see Paul Twomey. "Building on the Hamburg Statement and the G20 Roadmap for Digitalization: Toward a G20 framework for artificial intelligence in the workplace." At https://www.g20-insights.org/policy_briefs/building-on-the-hamburg-statement-and-the-g20-roadmap-for-digitalization-towards-a-g20-framework-for-artificial-intelligence-in-the-workplace/

[40] See https://www.un.org/en/content/digital-cooperation-roadmap/

[41] Regulation Of The European Parliament And Of The Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts, COM/2021/206 final. See https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206

The proposed EU rules would also prohibit "AI-based social scoring for general purposes done by public authorities," as well as AI systems that target "specific vulnerable groups" in ways that would "materially distort their behavior" to cause "psychological or physical harm." This could stop the use of AI for credit scoring, hiring, or some forms of surveillance advertising.

Like the European Commission's proposal, the Feasibility Study of CAHAI lays forth quite detailed provisions to address specific rights in European law. CAHAI is we are working on elaboration of elements of binding and non-binding instruments, and according to the timeline set by the Committee of Ministers (governing body composed by Foreign Ministers of 47 member countries) negotiations on a treaty are to start before May 2022.

In July 2021, representatives of Member States on UNESCO agreed on a draft recommendation on AI governance, to be submitted to the General Conference of UNESCO Member States in November 2021 for adoption[42]. The objectives of the draft Recommendation are:

(a) to provide a universal framework of values, principles and actions to guide States in the formulation of their legislation, policies or other instruments regarding AI;

(b) to guide the actions of individuals, groups, communities, institutions and private sector companies to ensure the embedding of ethics in all stages of the AI system life cycle;

(c) to promote respect for human dignity and gender equality, to safeguard the interests of present and future generations, and to protect human rights, fundamental freedoms, and the environment and ecosystems in all stages of the AI system life cycle;

(d) to foster multi-stakeholder, multidisciplinary and pluralistic dialogue about ethical issues relating to AI systems; and

(e) to promote equitable access to developments and knowledge in the field of AI and the sharing of benefits, with particular attention to the needs and contributions of LMICs, including LDCs, LLDCs and SIDS.

The draft recommendations include a call for an international regulatory framework to ensure that AI benefits humanity as a whole and respect, protection and promotion of human dignity, human rights and fundamental freedoms.

Questions considered in discussion among the Sub Committee

Q What can be achieved in an international agreement? How can we achieve consensus of policy across the three dominant models of data governance: Enterprise-centred Internet (US), State-centred Internet (China) and Citizen-centred Internet (the EU and other OECD partner countries)?

[42] See https://en.unesco.org/artificial-intelligence/ethics

Q What level of further detail above the G20 principles could be adopted by a broad range of states?

Q What does an international agreement accept or ignore in the EU draft legislation?

Q How does this effort relate to the meetings of governmental experts and officials in Geneva under the auspices of the Convention on Certain Conventional Weapons[43] to continue trying to find consensus on next steps in regulating the next class of automated weapons?

Conclusions of the Sub Committee's video conference discussions

The Subcommittee concluded that the ambition of the Policy Lab should be an international accord with the greatest appeal to all countries (at least members of the UN). In this sense it should be more like the Kyoto Agreement or existing UN human rights treaties rather than a treaty with a more limited likely membership such as the treaty on the prohibition of nuclear weapons.

As we examined the various statements and draft multilateral documents on AI we realized that we were not comparing apples with apples, but rather we were dealing with more vague apples and quite specific pears. The challenge for our first Sub-Committee discussion was what mix of the two do we think is realistic.

To that end the Subcommittee recommended that the definition of human rights being protected by the proposed international accord be as widely accepted as possible. We focused on those outlined by the 1948 Universal Declaration of Human Rights. These could be expanded to include rights outlined in the nine core UN human rights treaties:

- the International Covenant on Civil and Political Rights (ICCPR)
- the International Covenant on Economic, Social and Cultural Rights (ICESCR)
- the International Convention on the Elimination of All Forms of Racial Discrimination (CERD)
- the Convention on the Elimination of All Forms of Discrimination against Women (CEDAW)
- the Convention against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment (CAT)
- International Convention on the Protection of the Rights of All Migrant Workers and Members of Their Families (ICMW)
- the Convention on the Rights of the Child (CRC)
- the Convention on the Rights of Persons with Disabilities (CRPD).[44]

The Subcommittee discussion recognised that even these expanded treaties are controversial in some quarters and can be see as being deriving from the basic 30 principles outlined in the Universal Declaration. Hence our focus on the 1948 document. Indeed we concluded that 28 of the rights outlined could be negatively affected by AI, while 26 could be promoted through careful application of AI

[43] See https://www.un.org/disarmament/
[44] See https://www.ohchr.org/en/professionalinterest/pages/coreinstruments.aspx

applications. (See Appendix C). Further, we though that specific focus should also be made on limiting the use of AI to manipulate users, particularly the vulnerable.

We also considered that the move to an international accord should not be held hostage to the difficulty governments are having to find consensus on regulating the next class of autonomous weapons. To the degree it is possible the two discussions should not be linked.

As for what wording to move forward on for the negotiation of an international accord, the subcommittee discussion considered that the G20 statement should be considered a good starting position (perhaps augmented by some of the wording of the UNESCO recommendation if and how it is approved by the General Assembly). The members in discussion suspected that the EU draft legislation and even the CAHAI documents may fail to attract the universal approval we think is necessary for a global accord.

Finally, the discussion in the Sub Committee suggested that an international accord should also call for transparency and a call for pause and international review (if not a total moratorium) on the transition to General Intelligence by AI initiatives in their jurisdictions. An unrestricted move to General Intelligence would in our view pose a very significant potential threat to human rights.

Appendix A

G20 AI Principles

The G20 supports the Principles for responsible stewardship of Trustworthy AI in Section 1 and takes note of the Recommendations in Section 2.

Section 1: Principles for responsible stewardship of trustworthy AI

1.1. Inclusive growth, sustainable development and well-being
Stakeholders should proactively engage in responsible stewardship of trustworthy AI in pursuit of beneficial outcomes for people and the planet, such as augmenting human capabilities and enhancing creativity, advancing inclusion of underrepresented populations, reducing economic, social, gender and other inequalities, and protecting natural environments, thus invigorating inclusive growth, sustainable development and well-being.

1.2. Human-centered values and fairness
a) AI actors should respect the rule of law, human rights and democratic values, throughout the AI system lifecycle. These include freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, fairness, social justice, and internationally recognized labor rights.
b) To this end, AI actors should implement mechanisms and safeguards, such as capacity for human determination, that are appropriate to the context and consistent with the state of art.

1.3. Transparency and explainability
AI Actors should commit to transparency and responsible disclosure regarding AI systems. To this end, they should provide meaningful information, appropriate to the context, and consistent with the state of art:
i. to foster a general understanding of AI systems;
ii. to make stakeholders aware of their interactions with AI systems, including in the workplace;
iii. to enable those affected by an AI system to understand the outcome; and,
iv. to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision.

1.4. Robustness, security and safety
a) AI systems should be robust, secure and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety risk.
b) To this end, AI actors should ensure traceability, including in relation to datasets, processes and decisions made during the AI system lifecycle, to enable analysis of the AI system's outcomes and responses to inquiry, appropriate to the context and consistent with the state of art.
c) AI actors should, based on their roles, the context, and their ability to act, apply a systematic

risk management approach to each phase of the AI system lifecycle on a continuous basis to address risks related to AI systems, including privacy, digital security, safety and bias.

1.5. Accountability
AI actors should be accountable for the proper functioning of AI systems and for the respect of the above principles, based on their roles, the context, and consistent with the state of art.

Section 2: National policies and international co-operation for trustworthy AI

2.1. Investing in AI research and development
a) Governments should consider long-term public investment, and encourage private investment, in research and development, including inter-disciplinary efforts, to spur innovation in trustworthy AI that focus on challenging technical issues and on AI-related social, legal and ethical implications and policy issues.
b) Governments should also consider public investment and encourage private investment in open datasets that are representative and respect privacy and data protection to support an environment for AI research and development that is free of inappropriate bias and to improve interoperability and use of standards.

2.2. Fostering a digital ecosystem for AI
Governments should foster the development of, and access to, a digital ecosystem for trustworthy AI. Such an ecosystem includes in particular digital technologies and infrastructure, and mechanisms for sharing AI knowledge, as appropriate. In this regard, governments should consider promoting mechanisms, such as data trusts, to support the safe, fair, legal and ethical sharing of data.

2.3 Shaping an enabling policy environment for AI
a) Governments should promote a policy environment that supports an agile transition from the research and development stage to the deployment and operation stage for trustworthy AI systems. To this effect, they should consider using experimentation to provide a controlled environment in which AI systems can be tested, and scaled-up, as appropriate.
b) Governments should review and adapt, as appropriate, their policy and regulatory frameworks and assessment mechanisms as they apply to AI systems to encourage innovation and competition for trustworthy AI.

2.4. Building human capacity and preparing for labor market transformation
a) Governments should work closely with stakeholders to prepare for the transformation of the world of work and of society. They should empower people to effectively use and interact with AI systems across the breadth of applications, including by equipping them with the necessary skills.
b) Governments should take steps, including through social dialogue, to ensure a fair transition for workers as AI is deployed, such as through training programs along the working life, support for

those affected by displacement, and access to new opportunities in the labor market.

c) Governments should also work closely with stakeholders to promote the responsible use of AI at work, to enhance the safety of workers and the quality of jobs, to foster entrepreneurship and productivity, and aim to ensure that the benefits from AI are broadly and fairly shared.

2.5. International co-operation for trustworthy AI
a) Governments, including developing countries and with stakeholders, should actively cooperate to advance these principles and to progress on responsible stewardship of trustworthy AI.
b) Governments should work together in the OECD and other global and regional fora to foster the sharing of AI knowledge, as appropriate. They should encourage international, cross-sectoral and open multi-stakeholder initiatives to garner long-term expertise on AI.
c) Governments should promote the development of multi-stakeholder, consensus-driven global technical standards for interoperable and trustworthy AI.
d) Governments should also encourage the development, and their own use, of internationally comparable metrics to measure AI research, development and deployment, and gather the evidence base to assess progress in the implementation of these principles.

Appendix B

Europe fit for the Digital Age: Commission proposes new rules and actions for excellence and trust in Artificial Intelligence

Brussels, 21 April 2021

The Commission proposes today new rules and actions aiming to turn Europe into the global hub for trustworthy Artificial Intelligence (AI). The combination of the first-ever legal framework on AI and a new Coordinated Plan with Member States will guarantee the safety and fundamental rights of people and businesses, while strengthening AI uptake, investment and innovation across the EU. New rules on Machinery will complement this approach by adapting safety rules to increase users' trust in the new, versatile generation of products.

Margrethe **Vestager**, Executive Vice-President for a Europe fit for the Digital Age, said: *"On Artificial Intelligence, trust is a must, not a nice to have. With these landmark rules, the EU is spearheading the development of new global norms to make sure AI can be trusted. By setting the standards, we can pave the way to ethical technology worldwide and ensure that the EU remains competitive along the way. Future-proof and innovation-friendly, our rules will intervene where strictly needed: when the safety and fundamental rights of EU citizens are at stake."*

Commissioner for Internal Market Thierry **Breton** said: *"AI is a means, not an end. It has been around for decades but has reached new capacities fueled by computing power. This offers immense potential in areas as diverse as health, transport, energy, agriculture, tourism or cybersecurity. It also presents a number of risks. Today's proposals aim to strengthen Europe's position as a global hub of excellence in AI from the lab to the market, ensure that AI in Europe respects our values and rules, and harness the potential of AI for industrial use."*

The new **AI regulation** will make sure that Europeans can trust what AI has to offer. Proportionate and flexible rules will address the specific risks posed by AI systems and set the highest standard worldwide. The **Coordinated Plan** outlines the necessary policy changes and investment at Member States level to strengthen Europe's leading position in the development of human-centric, sustainable, secure, inclusive and trustworthy AI.

The European approach to trustworthy AI

The new rules will be applied directly in the same way across all Member States based on a future- proof definition of AI. They follow a risk-based approach:

Unacceptable risk: AI systems considered a clear threat to the safety, livelihoods and rights of people **will be banned**. This includes AI systems or applications that manipulate human behaviour to circumvent users' free will (e.g. toys using voice assistance encouraging dangerous behaviour of minors) and systems that allow 'social scoring' by governments.

High-risk: AI systems identified as high-risk include AI technology used in:

- **Critical infrastructures** (e.g. transport), that could put the life and health of citizens at risk;
- **Educational or vocational training**, that may determine the access to education and professional course of someone's life (e.g. scoring of exams);
- **Safety components of products** (e.g. AI application in robot-assisted surgery);
- **Employment, workers management and access to self-employment** (e.g. CV-sorting software for recruitment procedures);

- **Essential private and public services** (e.g. credit scoring denying citizens opportunity to obtain a loan);
- **Law enforcement** that may interfere with people's fundamental rights (e.g. evaluation of the reliability of evidence);
- **Migration, asylum and border control management** (e.g. verification of authenticity of travel documents);
- **Administration of justice** and **democratic processes** (e.g. applying the law to a concrete
- set of facts).
- High-risk AI systems will be subject to **strict obligations** before they can be put on the market:
- **Adequate risk assessment and mitigation systems;**
- **High quality of the datasets** feeding the system to minimise risks and discriminatory outcomes;
- **Logging of activity to ensure traceability of results**;
- **Detailed documentation** providing all information necessary on the system and its purpose for authorities to assess its compliance;
- **Clear and adequate information** to the user;
  Appropriate human oversight measures to minimise risk;
  High level of **robustness**, **security** and **accuracy**.

In particular, **all remote biometric identification** systems are considered high risk and subject to strict requirements. Their live use in publicly accessible spaces for law enforcement purposes is prohibited in principle. Narrow exceptions are strictly defined and regulated (such as where strictly necessary to search for a missing child, to prevent a specific and imminent terrorist threat or to detect, locate, identify or prosecute a perpetrator or suspect of a serious criminal offence). Such use is subject to authorisation by a judicial or other independent body and to appropriate limits in time, geographic reach and the data bases searched.

Limited risk, i.e. AI systems with specific transparency obligations: When using AI systems such as chatbots, users should be aware that they are interacting with a machine so they can take an informed decision to continue or step back.

Minimal risk: The legal proposal allows the free use of applications such as AI-enabled video games or spam filters. The vast majority of AI systems fall into this category. The draft Regulation does not intervene here, as these AI systems represent only minimal or no risk for citizens' rights or safety.

In terms of governance, the Commission proposes that national competent market surveillance authorities supervise the new rules, while the creation of a **European Artificial Intelligence Board** will facilitate their implementation, as well as drive the development of standards for AI. Additionally, voluntary codes of conduct are proposed for non-high-risk AI, as well as regulatory sandboxes to facilitate responsible innovation.

The European approach to excellence in AI

Coordination will strengthen Europe's leading position in human-centric, sustainable, secure, inclusive and trustworthy AI. To remain globally competitive, the Commission is committed to fostering innovation in AI technology development and use across all industries, in all Member States.

First published in 2018 to define actions and funding instruments for the development and uptake of AI, the **Coordinated Plan on AI** enabled a vibrant landscape of national strategies and EU funding for public-private partnerships

and research and innovation networks. The comprehensive update of the Coordinated Plan proposes concrete joint actions for collaboration to ensure all efforts are aligned with the European Strategy on AI and the European Green Deal, while taking into account new challenges brought by the coronavirus pandemic. It puts forward a vision to accelerate investments in AI, which can benefit the recovery. It also aims to spur the implementation of national AI strategies, remove fragmentation, and address global challenges.

The updated Coordinated Plan will use funding allocated through the **Digital Europe** and **Horizon Europe** programmes, as well as the **Recovery and Resilience Facility** that foresees a 20% digital expenditure target, and **Cohesion Policy** programmes, to:

- **Create enabling conditions for AI's development** and uptake through the exchange of policy insights, data sharing and investment in critical computing capacities;
- **Foster AI excellence** 'from the lab to the market' by setting up a public-private partnership, building and mobilising research, development and innovation capacities, and making testing and experimentation facilities as well as digital innovation hubs available to SMEs and public administrations;
- **Ensure that AI works for people** and is a force for good in society by being at the forefront of the development and deployment of trustworthy AI, nurturing talents and skills by supporting traineeships, doctoral networks and postdoctoral fellowships in digital areas, integrating Trust into AI policies and promoting the European vision of sustainable and

  trustworthy AI globally;
- **Build strategic leadership** in high-impact sectors and technologies including environment by focusing on AI's contribution to sustainable production, health by expanding the cross-border exchange of information, as well as the public sector, mobility, home affairs and agriculture, and Robotics.

The European approach to new machinery products

Machinery products cover an extensive range of consumer and professional products, from robots to lawnmowers, 3D printers, construction machines, industrial production lines. The Machinery Directive, replaced by the new Machinery Regulation, defined health and safety requirements for machinery. This new Machinery Regulation will ensure that the new generation of machinery guarantees the safety of users and consumers, and encourages innovation. While the AI Regulation will address the safety risks of AI systems, the new Machinery Regulation will ensure the safe integration of the AI system into the overall machinery. Businesses will need to perform only one single conformity assessment.

Additionally, the new Machinery Regulation will respond to the market needs by bringing greater legal clarity to the current provisions, simplifying the administrative burden and costs for companies by allowing digital formats for documentation and adapting conformity assessment fees for SMEs, while ensuring coherence with the EU legislative framework for products.

Next steps

The European Parliament and the Member States will need to adopt the Commission's proposals on a European approach for Artificial Intelligence and on Machinery Products in the ordinary legislative procedure. Once adopted, the Regulations will be directly applicable across the EU. In parallel, the Commission will continue to collaborate with Member States to implement the actions announced in the Coordinated Plan.

Background

For years, the Commission has been facilitating and enhancing cooperation on AI across the EU to boost its competitiveness and ensure trust based on EU values.

Following the publication of the European Strategy on AI in 2018 and after extensive stakeholder consultation, the High-Level Expert Group on Artificial Intelligence (HLEG) developed Guidelines for Trustworthy AI in 2019, and an Assessment List for Trustworthy AI in 2020. In parallel, the first Coordinated Plan on AI was published in December 2018 as a joint commitment with Member States.

The Commission's White Paper on AI, published in 2020, set out a clear vision for AI in Europe: an ecosystem of excellence and trust, setting the scene for today's proposal. The public consultation on the White Paper on AI elicited widespread participation from across the world. The White Paper was accompanied by a 'Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics' concluding that the current product safety legislation contains a number of gaps that needed to be addressed, notably in the Machinery Directive.

For More Information

New rules for Artificial Intelligence – Questions and Answers New rules for Artificial Intelligence – Facts page

Communication on Fostering a European approach to Artificial Intelligence Regulation on a European approach for Artificial Intelligence

New Coordinated Plan on Artificial Intelligence Regulation on Machinery Products

EU-funded AI projects

Appendix C

List of 30 basic human rights and their possible impact by AI

The Universal Declaration of Human Rights was approved by the United Nations General Assembly at the Palais de Chaillot in Paris, France on 10 December 1948. Of the then 58 members of the United Nations, 48 voted in favor, none against, eight abstained, and two did not vote.

This declaration consists of 30 articles affirming an individual's rights.

Beside each right I have placed my judgement as to how a right could be affected by AI applications. You will see that the majority could be both negatively impacted and also positively impacted – a not unusual consequence of such a powerful enabling technology.

Paul Twomey
1 September 2021

1. All human beings are free and equal Negatively

All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.

2. No discrimination Both Negatively or Positively

Everyone is entitled to all the rights and freedoms, without distinction of any kind, such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status. Furthermore, no distinction shall be made on the basis of the political, jurisdictional or international status of the country or territory to which a person belongs.

3. Right to life Both Negatively or Positively

Everyone has the right to life, liberty and security of person.

4. No slavery Both Negatively or Positively

No one shall be held in slavery or servitude; slavery and the slave trade shall be prohibited in all their forms.

5. No torture and inhuman treatment Both Negatively or Positively

No one shall be subjected to torture or to cruel, inhuman or degrading treatment or punishment.

6. Same right to use law Both Negatively or Positively

Everyone has the right to recognition everywhere as a person before the law.

7. Equal before the law Both Negatively or Positively

All are equal before the law and are entitled without any discrimination to equal protection of the law. All are entitled to equal protection against any discrimination in violation and against any incitement to such discrimination.

8. Right to treated fair by court Both Negatively or Positively

Everyone has the right to an effective remedy by the competent national tribunals for acts violating the fundamental rights granted him by the constitution or by law.

9. No unfair detainment Both Negatively or Positively

No one shall be subjected to arbitrary arrest, detention or exile.

10. Right to trial Both Negatively or Positively

Everyone is entitled in full equality to a fair and public hearing by an independent and impartial tribunal, in the determination of his rights and obligations and of any criminal charge against him.

11. Innocent until proved guilty Negatively

Everyone charged with a penal offence has the right to be presumed innocent until proved guilty according to law in a public trial at which he has had all the guarantees necessary for his defence. No one shall be held guilty of any penal offence on account of any act or omission which did not constitute a penal offence, under national or international law, at the time when it was committed.

12. Right to privacy Both Negatively or Positively

No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks.

13. Freedom to movement and residence Both Negatively or Positively

Everyone has the right to freedom of movement and residence within the borders of each state. Everyone has the right to leave any country, including his own, and to return to his country.

14. Right to asylum Both Negatively or Positively

Everyone has the right to seek and to enjoy in other countries asylum from persecution. This right may not be invoked in the case of prosecutions genuinely arising from non-political crimes or from acts contrary to the purposes and principles of the United Nations.

15. Right to nationality Both Negatively or Positively

Everyone has the right to a nationality. No one shall be arbitrarily deprived of his nationality nor denied the right to change his nationality

16. Rights to marry and have family Both Negatively or Positively

Men and women of full age, without any limitation due to race, nationality or religion, have the right to marry and to found a family. They are entitled to equal rights as to marriage, during marriage and at its dissolution. Marriage shall be entered into only with the free and full consent of the intending spouses. The family is the natural and fundamental group unit of society and is entitled to protection by society and the State.

17. Right to own things Both Negatively or Positively

Everyone has the right to own property alone as well as in association with others. No one shall be arbitrarily deprived of his property.

18. Freedom of thought and religion Both Negatively or Positively

Everyone has the right to freedom of thought, conscience and religion; this right includes freedom to change his religion or belief, and freedom, either alone or in community with others and in public or private, to manifest his religion or belief in teaching, practice, worship and observance.

19. Freedom of opinion and expression Both Negatively or Positively

Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers.

20. Right to assemble Both Negatively or Positively

Everyone has the right to freedom of peaceful assembly and association. No one may be compelled to belong to an association.

21. Right to democracy Both Negatively or Positively

Everyone has the right to take part in the government of his country, directly or through freely chosen representatives. Everyone has the right of equal access to public service in his country.

22. Right to social security Both Negatively or Positively

Everyone, as a member of society, has the right to social security and is entitled to realization, through national effort and international co-operation and in accordance with the organization and resources of each State, of the economic, social and cultural rights indispensable for his dignity and the free development of his personality.

23. Right to work Both Negatively or Positively

Everyone has the right to work, to free choice of employment, to just and favourable conditions of work and to protection against unemployment. Everyone, without any discrimination, has the right to equal pay for equal work. Everyone has the right to form and to join trade unions for the protection of his interests.

24. Right to rest and holiday Both Negatively or Positively

Everyone has the right to rest and leisure, including reasonable limitation of working hours and periodic holidays with pay.

25. Right of social service Both Negatively or Positively

Everyone has the right to a standard of living adequate for the health and well-being of himself and of his family, including food, clothing, housing and medical care and necessary social services, and the right to security in the event of unemployment, sickness, disability, widowhood, old age or other lack of livelihood in circumstances beyond his control. Motherhood and childhood are entitled to special care and assistance. All children shall enjoy the same social protection.

26. Right to education Both Negatively or Positively

Everyone has the right to education. Education shall be free, at least in the elementary and fundamental stages. Elementary education shall be compulsory. Technical and professional education shall be made generally available and higher education shall be equally accessible to all on the basis of merit.

27. Right of cultural and art Both Negatively or Positively

Everyone has the right freely to participate in the cultural life of the community, to enjoy the arts and to share in scientific advancement and its benefits. Everyone has the right to the protection of the moral and material interests resulting from any scientific, literary or artistic production of which he is the author.

28. Freedom around the world Both Negatively or Positively

Everyone is entitled to a social and international order in which the rights and freedoms set forth in this Declaration can be fully realized.

29. Subject to law Both Negatively or Positively

Everyone has duties to the community in which alone the free and full development of his personality is possible. In the exercise of his rights and freedoms, everyone shall be subject only to such limitations as are determined by law solely for the purpose of securing due recognition and respect for the rights and freedoms of others and of meeting the just requirements of morality, public order and the general welfare in a democratic society.

30. Human rights can't be taken away Negatively

Nothing in this Declaration may be interpreted as implying for any State, group or person any right to engage in any activity or to perform any act aimed at the destruction of any of the rights and freedoms set forth herein.

Works Cited

Acquisti, Allesandro, and Christina M. Fong. 2015. "An Experiment in Hiring Discrimination via Online Social Networks." https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2031979.

Angwin, Julia, Jeff Larson, Surya Mattu and Lauren Kirchner. 2016. "Machine Bias." *ProPublica*, May 23. www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Balkin, Jack M. 2015. "Information Fiduciaries and the First Amendment." *UCDL Rev.* 49: 1183.

Barghava, Rahul. 2017. "The Algorithms Aren't Biased, We Are." *MIT Media Lab* (blog), January 3. https://medium.com/mit-media-lab/the-algorithms-arent-biased-we-are-a691f5f6f6f2.

Barocas, Solon, and Andrew D. Selbst. 2016. "Big Data's Disparate Impact." *California Law Review* 104: 671–732.

British Columbia First Nations Data Governance Initiative. 2017. *Decolonizing Data: Indigenous Data Sovereignty Primer*. April.

Burt, Andrew. 2017. "Is there a 'right to explanation' for machine learning in the GDPR?" IAPP *Privacy Tech* (blog), June 2. International Association of Privacy Professionals. https://iapp.org/news/a/is-there-a-right-to-explanation-for-machine-learning-in-the-gdpr/.

Cellan-Jones, Rory. 2014. "Stephen Hawking warns artificial intelligence could end mankind." BBC News, December 2.. www.bbc.com/news/technology-30290540.

Clifford, Catherine . 2017. "Hundreds of A.I. experts echo Elon Mush, Stephen Hawking in call for a ban on killer robots." CNBC, November 8. https://www.cnbc.com/2017/11/08/ai-experts-join-elon-musk-stephen-hawking-call-for-killer-robot-ban.html.

Englehardt, Steven, and Arvind Narayanan. 2016. "Online Tracking: A 1-Million-Site Measurement and Analysis." http://randomwalker.info/publications/OpenWPM_1_million_site_tracking_measurement.pdf.

Eubanks, Virgina. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* New York, NY: St. Martin's Press.

Executive Office of the President. 2014. *Big Data: Seizing Opportunities, Preserving Values*. Washington, DC: The White House. https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf.

Frey, Carl Benedikt and Michael A. Osborne. 2013. *The Future of Employment: How Susceptible Are Jobs to Computerisation?* September 17. Oxford, UK: Oxford Martin Programme on Technology and Employment. www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf.

G20. 2017a. *G20 Digital Economy Ministerial Conference. Düsseldorf 6-7 April 2017.* Declaration of the Ministers Responsible for the Digital Economy. Federal Ministry for Economic Affairs and Energy. www.bmwi.de/Redaktion/DE/Downloads/G/g20-digital-economy-ministerial-declaration-english-version.pdf?__blob=publicationFile&v=12.

G20. 2017b. "A ROADMAP for Digitalisation: Policies for a Digital Future." Annex paper 1 to the Declaration of the Ministers responsible for the Digital Economy. In *G20 Digital Economy Ministerial Conference. Düsseldorf 6-7 April 2017*, 10–15. www.bmwi.de/Redaktion/DE/Downloads/G/g20-digital-economy-ministerial-declaration-english-version.pdf?__blob=publicationFile&v=12.

G20. 2017c. "G20 Leaders' Declaration: Shaping an interconnected world." G20 Germany 2017 meetings, Hamburg, July 7-8. www.g20germany.de/Content/EN/_Anlagen/G20/G20-leaders-declaration.pdf;jsessionid=0C08AA235271BF43ECBB08BA059EE5B7.s6t2?__blob=publicationFile&v=11.

Gangadharan, Seeta P., Virginia Eubanks and Solon Barocas, eds. 2014. *Data and Discrimination: Collected Essays*. Washington, DC: New America. www.newamerica.org/oti/policy-papers/data-and-discrimination/.

Global Commission on Internet Governance. 2016. *One Internet: Final Report of the Global Commission on Internet Governance.* Waterloo, ON: CIGI.. www.cigionline.org/publications/one-internet.

Hurley, Dan. 2018. "Can an Algorithm Tell When Kids Are in Danger?" *New York Times*, January 2. www.nytimes.com/2018/01/02/magazine/can-an-algorithm-tell-when-kids-are-in-danger.html.

Jaume-Palasí, Lorena and Matthias Spielkamp. 2017. "Ethics and algorithmic processes for decision making and decision support." AlgorithmWatch Working Paper No. 2, June 1. Berlin, Germany: AlgorithmWatch. https://algorithmwatch.org/en/ethics-and-algorithmic-processes-for-decision-making-and-decision-support/.

Kirchner, Lauren. 2017. "New York City moves to create accountability for algorithms." *Ars Technica*, December 19. https://arstechnica.com/tech-policy/2017/12/new-york-city-moves-to-create-accountability-for-algorithms/.

KPMG International. 2016. "Rise of the humans: The integration of digital and human labor." KPMG International Cooperative, November. https://assets.kpmg.com/content/dam/kpmg/xx/pdf/2016/11/rise-of-the-humans.pdf.

Kroll, Joshua A., Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson and Harlan Yu. 2017. "Accountable algorithms." *University of Pennsylvania Law Review* 165: 633–705.

Lohr, Steve. 2017. "A.I. will transform the Economy. But How Much, and How Soon?" *New York Times*, November 30. www.nytimes.com/2017/11/30/technology/ai-will-transform-the-economy-but-how-much-and-how-soon.html.

Madden, Mary, Michele Gilman, Karen Levy and Alice Marwick. 2017. "Privacy, Poverty, and Big Data: A Matrix of Vulnerabilities for Poor Americans." *Washington University Law Review* 95 (1): 53–125.

Manyika, James, Susan Lund, Michael Chui, Jacques Bughin, Jonathan Woetzel, Parul Batra, Ryan Ko and Saurabh Sanghvi. 2017. *Jobs Lost, Jobs Gained: Workforce Transitions in a Time of Automation*. San Francisco, CA: McKinsey Global Institute. www.mckinsey.com/featured-insights/future-of-organizations-and-work/jobs-lost-jobs-gained-what-the-future-of-work-will-mean-for-jobs-skills-and-wages.

McGeveran, William. 2018. "The Duty of Data Security." *Minn. L. Rev.* 103: 1135.
Scholz, Lauren Henry. 2019. "Privacy Remedies." *Indiana Law Journal*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3159746

Microsoft. 2018. *The Future Computed: Artificial Intelligence and its role in society*. Redmond, WA: Microsoft. https://blogs.microsoft.com/uploads/2018/02/The-Future-Computed_2.8.18.pdf.

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York, NY: New York University Press.

Obar, Jonathan A. and Anne Oeldorf-Hirsch. 2016. "The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services." https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2757465.

Obar, Jonathan, and Brenda McPhail. 2018. "Preventing Big Data Discrimination in Canada: Addressing Design, Consent and Sovereignty Challenges." In *Data Governance in the Digital Age*: *Special Report*, 56–64. Waterloo, ON: CIGI. https://www.cigionline.org/publications/data-governance-digital-age.

O'Neil, Cathy. 2017. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York, NY: Broadway Books.

Ovenden, James. 2016. "AI In Developing Countries: Artificial intelligence isn't just for self driving cars." Innovation Enterprise, October 6. https://channels.theinnovationenterprise.com/articles/ai-in-developing-countries.

Pasquale, Frank. 2015. *The Black Box Society: The Secret Algorithms that Control Money and Information.* Cambridge, MA: Harvard University Press.

Purdy, Mark, and Paul Daugherty. 2016. "Why Artificial Intelligence Is the Future of Growth." Accenture Institute for High Performance, September 28. www.accenture.com/us-en/insight-artificial-intelligence-future-growth.

Reidenberg, Joel R., Travis Breaux, Lorrie Faith Cranor, Brian French, Amanda Grannis, James T. Graves, Fei Liu, Aleecia McDonald, Thomas B. Norton, Rohan Ramanath, N. Cameron Russell, Norman Sadeh and Florian Schaub. 2015. "Disagreeable Privacy Policies: Mismatches Between Meaning and Users'

Understanding." *Berkeley Technology Law Journal* 30 (1): 39–68. https://scholarship.law.berkeley.edu/cgi/viewcontent.cgi?article=2053&context=btlj.

Sandvig, Christian, Kevin Hamilton, Karrie Karahalios and Cedric Langbort. 2016. "When the Algorithm Itself Is a Racist: Diagnosing Ethical Harm in the Basic Components of Software." *International Journal of Communication* 10: 4972–90. http://social.cs.uiuc.edu/papers/pdfs/Sandvig-IJoC.pdf.

Scannell, R. Joshua. 2016. "Broken Windows, Broken Code." Reallifemag.com, August 29. http://reallifemag.com/broken-windows-broken-code/.

Solove, Daniel J. 2013. "Introduction: Privacy Self-Management and the Consent Dilemma." *Harvard Law Review* 126: 1880–1903. https://pdfs.semanticscholar.org/809c/bef85855e4c5333af40740fe532ac4b496d2.pdf.

Susser, Daniel, Beate Roessler, and Helen Nissenbaum. 2019. "Technology, Autonomy, and Manipulation." *Internet Policy Review* 8 (2).

The Economist. 2018. "AI-Spy: The workplace of the future." March 28. https://www.economist.com/leaders/2018/03/28/the-workplace-of-the-future.

Turow, Joseph. 2011. *The Daily You: How the New Advertising Industry Is Defining Your Identity and Your Worth*. New Haven, CT: Yale University Press.

UNI Global Union. 2017. *Top 10 Principles for Ethical Artificial Intelligence*. Nyon, Switzerland: UNI Global Union. www.thefutureworldofwork.org/opinions/10-principles-for-ethical-ai/.UNI Global Union. 2018. *Top 10 Principles for Workers' Data Privacy and Protection*. Nyon, Switzerland. www.thefutureworldofwork.org/docs/10-principles-for-workers-data-rights-and-privacy/.

Wachter, Sandra, and Brent Mittelstadt. 2019. "A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI." *Columbia Business Law Review*

Wexler, Rebecca. 2017. "When a Computer Program Keeps You in Jail." New York Times, June 13. www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html.

[i] Trewin, S., Basson, S., Muller, M., Branham, S., Treviranus, J., Gruen, D., Hebert, D., Lyckowski, N. and Manser, E., 2019. Considerations for AI fairness for people with disabilities. AI Matters, 5(3), pp.40-63.

Policy Lab:

Fundamental Rights in AI Digital Societies: Towards an International Accord

Sub-Committee 2:

Transatlantic approaches to protect fundamental rights in AI & digital spaces

Issues Paper (23/08/2021)

Véronique Choquette, Senior Policy Advisor, Club de Madrid
Jerry Jones, EVP, Chief Legal & Ethics Officer, LiveRamp

1. Introduction

Artificial intelligence (AI) is a disruptive technology, with profound implications in individual life, in society and in geopolitics. It has given us new tools for daily life, from smart utilities to virtual assistants, and revolutionized how we - citizens, businesses, governments and others – relate to each other. Our information ecosystem, our economic relations and our decision-making systems, from credit ratings to school admissions and public resource allocation, are all increasingly driven by algorithms. And this transformation is altering the global balance of power, changing the factors that drive economic growth and occasioning growing tensions around technological innovation, data collection and governance, and the relationship between citizens and digital technologies.

As any major societal transformation, digital transformation is bringing about new threats and opportunities for fundamental rights. It is opening up new channels of expression, association and citizen engagement in public affairs, and new tools for institutional transparency and accountability. Assistive technologies based on AI and robotics can help large groups of citizens fulfil their rights to safety, autonomy and dignity, while AI-assisted decision-making can improve the quality and efficiency of public service delivery, from education to healthcare and security services.

But digital transformation is also posing new threats to fundamental rights. It is concentrating economic and political power in the hands of giant corporate actors; providing human rights offenders, including repressive governments, with new tools for censorship, monitoring and crack-down; and opening new pathways for foreign surveillance, interference and warfare. The use of AI-assisted decision-making tools is also jeopardizing the crucial role of human judgement and raising questions of accountability for all sorts of decisions, in business, public service and military areas alike.

There is no question that legal and governance frameworks must evolve in order to ensure that the continued development and deployment of AI and digital technologies protects rather than threatens fundamental rights. And since technology advances faster than law-making, at least in deliberative democratic systems, it is also clear that the governance framework on AI and digital technologies must provide forward-looking guardrails, protecting fundamental rights in the face of future, as well as current, technology.

There is also broad agreement around the need for international coordination on this score. The UN Secretary-General's Roadmap for Digital Cooperation calls for "supporting global cooperation on artificial intelligence that is trustworthy, human-rights based, safe and sustainable and promotes peace". Yet the locus of global leadership to protect fundamental rights in the face of digital technologies remains unclear. How can the global community reach agreement on a basic set of fundamental rules to guide future technology development? And how should such an agreement, hypothetical as it might be, be enforced?

The strategic approaches of China and Russia to develop and deploy technologies unhindered by human rights considerations undoubtedly leave the world's democratic powerhouses – the EU and the US -- on the same side of the issue. For this reason, it has been argued that a Transatlantic alliance could be a natural starting point for a global accord on AI governance aimed at protecting fundamental rights. But it does not follow that Transatlantic cooperation on AI governance is a straightforward enterprise. Could the EU and the US, together, lay the groundwork for a global agreement on a basic set of fundamental rules to guide AI technology development?

The aim of this paper is to present the EU and the US' approaches to AI and digital technologies, with a view to gauging the possibility for joint EU-US leadership towards a global accord to protect fundamental rights in the AI and digital spaces. It is one of three papers that will feed into the discussions hosted by Club de Madrid and the Boston Global Forum at the Policy Lab on *Fundamental Rights in AI Digital Societies: Towards an International Accord* on 7-9 September 2021.

2. The EU's rights-oriented approach to AI governance

The EU is neither a global leader in digital technology innovation, nor a quick adopter of large-scale AI applications. Despite recent efforts to stimulate investment in AI, its tech industry continues to trail behind US and Chinese innovators, marred by the scarcity of private funding, the lack of a European hub for AI expertise, severe brain drain, low appetite for AI solutions in the public sector and the relatively limited availability of data to feed AI solutions under the EU's General Data Protection Regulation. Regulatory fragmentation, in the absence of a complete Digital Single Market, also limits the possibilities for European innovators to scale up.

The relatively slow development of home-grown digital technologies, combined with the centrality of human rights protection in the European project, have led the EU to approach AI primarily as a rights issue. Stimulating the digital industry while ensuring that the deployment of AI technologies does not hinder citizens' rights has become the central axis of the EU's AI policy; and individual data ownership, wherein data belongs to the individual that produces it rather than the company that harvests it, its basic tenet. Sometimes portrayed as the "third way" between so-called American technological libertarianism and Chinese technological authoritarianism, the EU aims to lead in the rights-based, ethical governance of AI technologies.

Early impulses for AI regulation came from the European Parliament during the Juncker Commission (2014-2019). The 2018 Communication on Artificial Intelligence for Europe laid out the Commission's first approach to AI, articulated around investment promotion, socio-economic change and ethics. Two years later, a High-Level Expert Group on AI advised the Commission on necessary policy and regulatory changes, leading the von der Leyen Commission (2019-2024) to make *Europe fit for the digital age* one of its priorities, under the leadership of Executive Vice-President Margrethe Vestager. Building on the rights-based approach underpinning its now rolled-out General Data Protection Regulation, the Commission issued in early 2021 a proposal for harmonized rules on AI in the EU – the Communication on Fostering a European Approach to AI, or EU AI Act.

The EU AI Act is still far from becoming law, and has received heavy criticism from technology enthusiasts and human rights defenders alike. The former have called it a regulatory straightjacket that will stifle innovation, while the latter lament that it does not go far enough to protect the rights of end-users, that is, citizens. Nevertheless, the EU AI Act is laudable as the world's first attempt at a comprehensive rights-based AI regulation, and relevant as an illustration of the EU's approach to AI governance. It sets out a three-tiered regulatory structure that would ban some uses of AI altogether (such as social scoring and indiscriminate surveillance), heavily regulate high-risk uses, and lightly regulate less risky AI systems -- complete with *ex ante* conformity assessments and the creation of light monitoring structures. While no other jurisdiction in the world has a similar scheme in place for AI, White House National Security Advisor Jake Sullivan pointed out that it bears similarity with systems used by US financial regulators.

The EU AI Act, like the policy reflections that preceded it, is comprehensive in its treatment of AI as a domestic fundamental rights issue. While putting rights first, it also seeks to enable, through greater market integration and regulatory certainty, the role of AI as a motor of future economic growth in the region. However, it entirely leaves aside the military uses of AI, as well as any discussion of the associated considerations related to strategic interests and geopolitics. EU institutions, who are naturally shy of military matters for which they lack competencies, are not entirely to blame for this omission. Since 2018, the EU has been encouraging its Member States to adopt national AI strategies, as part of the Commission's Coordinated Plan on AI. Of the 21 Member States who have adopted or drafted strategies so

far, only France and the Netherlands – two exporters of AI-based military technology -- touch upon the geopolitical implications of AI, pointing out the strategic importance of the EU's digital autonomy. For all others, AI remains a fundamentally domestic issue.

Yet there have been calls for the EU to engage with the geopolitical dimension of AI, and for EU leadership for the governance of AI in the military space. Continuing to disregard the implications of AI for its foreign relations and geopolitical influence, warns the European Council on Foreign Relations, would lock the EU into the role of mere mediator between the real technological powers, the US and China.

The European Parliament has also called for human dignity and human rights to be respected in all EU defence-related activities, including those involving AI systems; and it has expressed its support for a ban on lethal autonomous weapon systems (LAWS), also known as killer robots. The European Defence Agency – an EU Agency mandated to promote collaboration among EU Member States on defence matters -- has been working since 2016 on plans for EU collaboration on AI in defence, but results have been slow to come, an indicator of the difficulty of EU leadership in military matters.

3. The US' military-strategic approach to AI governance

While the term *technological libertarianism* exaggerates and oversimplifies the US' approach to AI governance, it is undeniable that, compared to the EU, the US has been approaching AI with a lighter regulatory foot. Protective of the global leadership of the country's tech sector, and undesiring to risk muffling innovation with red tape, both the Obama and Trump administrations have stayed away from comprehensive regulation on AI. The cession of data ownership by individuals to tech companies through informed consent – such as that given in User Agreements – has been deemed legitimate, and tech companies have been encouraged to adopt voluntary standards of responsibility in the use of such data. Adjusting existing regulatory frameworks with the minimum touches necessary to address known risks, has been the preferred approach to AI government regulation.

Federal guidance on AI ethics is not entirely absent – the White House Office of Management and Budget released in 2020 a set of policy principles for regulating AI articulated around the objective to promote innovation while protecting privacy, civil rights and American values – but the most ambitious regulations on AI in the country have come from local and state administrations. Federal efforts, such as Obama's twin reports on Preparing for the Future of AI and National AI R&D Strategic Plan (2016), Trump's American AI Initiative (2019) or the recent announcement by the Biden administration of a National Artificial Intelligence (AI) Research Resource Task Force (2021), emphasize the strategic importance of AI and AI innovation for the US economy and security; and while they do mention the implications of AI for human rights, they fall short of suggesting that regulation is the solution. The creation in 2019 by the US Congress of the National Security Commission on Artificial Intelligence (NSCAI)

confirms that the focus on the strategic dimension of AI is not an Executive feat – it is the American approach.

Viewing AI as a primarily strategic issue, it is only natural that the US should trail behind the EU on the governance of AI as a domestic fundamental rights issue, but lead on governing its applications in the strategic sphere. In 2020, the US Department of Defense adopted a series of ethical principles for the design, deployment and adoption of military applications of AI, becoming the first US public administration to prescribe an AI norm that goes much further than corporate voluntary standards. The principles establish *inter alia* that human beings must remain responsible for the development, deployment, use and outcomes of AI systems; algorithms used in combat must avoid unintended bias; and AI systems must be programmed to stop themselves if they see that they might be causing problems.

While there have been calls for the US to lead the development of joint military standards on AI, not least through NATO, it has so far been choosing its partners carefully. NSCAI recommended the Five Eyes Alliance (US, UK, Canada, Australia and New Zealand) as a first locus of collaboration, and in 2020 the Pentagon expanded its consultations to a group of 13 countries through the AI Partnership for Defence (Australia, Canada, Denmark, Estonia, Finland, France, Israel, Japan, Norway, the Republic of Korea, Sweden, and the United Kingdom).

Beyond this normative effort, the US is also making it a priority to leverage AI to strengthen its military capacity, and that of its allies, through improved systems for asset protection and information processing. This has been encouraged by NSCAI as an essential measure to preserve national security and remain competitive with China and Russia. AI safety in military operations – protecting US military AI systems from foreign interference – is next on the list of priorities. The ultimate objective of these efforts is obviously to build a countervailing force against China, who is also seen to be investing heavily in new technologies and implementing them in new advanced weapon systems, without – it is suspected – the kind of ethical considerations to which the US has yet remained committed.

4. Transatlantic cooperation: Where to start?

The two different AI approaches put forward by the EU and the US -- with the former focused on the domestic socio-economic implications of AI and the latter on leveraging technology to preserve and strengthen its geopolitical power – appear to be rather complementary than incompatible. While the US does not share the EU's appetite for comprehensive regulation, and the EU has neither the competence nor the strategic unity to match US leadership in the military space, their different strategic objectives are not conflicting and rest on shared values. Their different approaches to data ownership, however, wherein the EU seeks to give individuals full control over their data and how it is used, while the US allows the unfettered cession of data rights to private companies, limits the scope for agreement on what guardrails are needed to guide the development of future digital and AI technologies.

In December 2020, the European Commission greeted the incoming Biden administration with an ambitious blueprint for Transatlantic cooperation (New Transatlantic Agenda for Global Change), including two proposals related to digital technologies: the creation of a Trade and Technology Council, and working together on global standards for AI governance. Leaders on both sides officially established the EU-US Trade and Technology Council at the EU-US Summit of June 2021, stating among its goals "to cooperate on compatible and international standards development; to facilitate regulatory policy and enforcement cooperation and, where possible, convergence; […] and to feed into coordination in multilateral bodies and wider efforts with like-minded partners, with the aim of promoting a democratic model of digital governance." The Council will operate through working groups, whose initial agendas will focus on technology standards cooperation, including on AI, data governance and the misuse of technology threatening security and human rights.

If the US-UK Science and Technology Agreement of 2017 is any precedent, there are reasons to hope that the EU-US Trade and Technology Council could serve not only to reach agreements in areas where interests align, but also to build enough goodwill to open discussions on divergent issues. Regulatory changes to the business environment surrounding AI is one area in which the EU and the US could see eye to eye quickly. Market concentration in the data economy is testing the limits of anti-trust laws on both sides of the Atlantic, and the creation of a EU-US Joint Technology Competition Policy Dialogue, alongside the Trade and Technology Council, shows a willingness to cooperate in the quest for solutions.

4.1. Cooperation on AI regulation

Despite the willingness expressed at the EU-US Summit and engrained in the mandate of the EU-US Trade and Technology Council to cooperate on technology standards for ethical and trustworthy AI, the EU's appetite for comprehensive regulation will in all likelihood continue to meet with opposition from US business interests. But there is scope for cooperation around the shared objective to provide companies with regulatory stability and administrative facility.

The concept of *high-risk uses of AI* is a central element in the proposed EU AI Act; only high-risk AI systems would be subject to the toughest restrictions and controls. Agreeing with the US on a common definition of high-risk AI, even if subject to different frameworks on either side of the Atlantic, would provide more clarity for companies operating in the two regions, and lay the foundation for cooperation on the governance – through regulation or other instruments - of high-risk applications.

Easing the administrative burden on companies by arranging for the mutual recognition of certification schemes is another objective around which EU and US interests could meet. In the (likely) event that the EU moves first with a comprehensive AI regulation, an arrangement to allow US companies to obtain certification through the US government could help set basic standards accepted on both sides and facilitate inter-operability. Mutual recognition agreements could also be built piece-by-piece, through bilateral consultations between

specialized agencies, who are often responsible for technical norms in the US, with support from the new Trade and Technology Council.

4.2. Cooperation on AI geopolitics

While cooperation on the geopolitical implications of AI was not explicitly mentioned at the EU-US Summit last June, there are some encouraging signs that closer collaboration on that score might be in the cards for the near future. The Summit declaration refers to new arrangements for closer partnership in security and defence, such as US participation in an EU military mobility project and closer engagement with the European Defence Agency. It also includes a commitment to cooperate on "the full range of issues" in their relationship with China.

On the EU's side, there are also early signs that awareness of the geopolitical implications of AI is beginning to take root. The concept of digital and technological sovereignty has appeared in the conversations on the Future of Europe; the European External Action Service has started regarding technology, connectivity and data flows as a key dimension of the EU's external relations; and the European Council has called for a geostrategic and global approach to connectivity. In public interventions calling for the protection of fundamental rights in the digital space, the European Commission has also started referring specifically to China as a source of concern in its own territory and globally. This bodes well for a growing willingness from the EU to engage with the US' geostrategic approach to AI.

Should the EU and the US wish to make a common front against China's AI advances – whether for ethical or for geopolitical reasons -- US researchers have put forward a number of commercial strategies that would not require military competence yet would make a huge strategic difference. This could include, for example, coordinating investment screening procedures and establishing common export controls on key supply chain components going into the Chinese AI industry.

There are also many opportunities for cooperation on military uses of AI, that would both promote an ethical approach and strengthen mutual capacity. EU-US cooperation on the ethical use (or ban) of killer robots and other combat-related AI systems appears highly unlikely – but cooperation on non-controversial uses of AI in military services such as logistics, financial management, personnel services, and health care could help bring allies closer together, establish joint procedures, and ensure interoperability.

Of course, NATO provides an additional – and some have said an ideal -- forum to bring together EU and US approaches to the geopolitical dimension of AI and set standards for military AI. But the same barriers that have hindered both EU engagement and EU-US cooperation in these areas apply with equal force within NATO. Its members' widely divergent priorities make consensus unlikely on key issues, including those – like a ban on killer robots –

that seem most obvious to human rights defenders and civil society organizations across the alliance's territory.

NATO's AI strategy – in the works since 2019 and slated to be released sometime soon - is expected to identify ways to operate AI defence systems responsibly, identify military applications for AI, and set up joint AI testing facilities. It should also set ethical guidelines around the governance of AI systems, with a focus on human control over and accountability for the actions of AI systems.

5. From Transatlantic to global

In sum, despite fundamentally different approaches to AI, there appear to be a number of promising avenues for greater Transatlantic cooperation around the governance of AI-based technologies, whether in the domestic sphere to protect fundamental rights, or in the geopolitical sphere around common national and global security interests. The Summit for Democracies, convened by President Biden for 9-10 December 2021, may provide an additional forum where leaders from both sides of the Atlantic may reinforce their common commitment to fundamental rights, including in the digital sphere. The possibility for the EU and the US to see eye-to-eye on the full range of issues pertaining to the development of a common set of basic rules to guide the development of AI technologies, however, remains limited by the different values that each region is strategically choosing to prioritize.

Going back to the initial premise of this paper, namely the quest for a locus of global leadership for the rights-based governance of digital technologies, it would appear that a Transatlantic alliance, even if it were reached, with the limitations imposed by the different approaches put forward by the EU and the US, may not have enough horsepower to pull the train.

The US' interest in the Transatlantic relationship has been waning, as its strategic considerations have been turning increasingly towards the Pacific. Both the EU and the US are emphasizing the importance of working with other actors, as well as with each other, on AI issues. The US' AI Partnership for Defence and the Global Partnership for AI (initially spearheaded by Canada and France, and now housed at the OECD) are two examples of collaborative structures that aim to bring a broader group of like-minded partners in the conversation on AI governance. A number of global organizations, including UNESCO, are also beginning to weigh in with specific initiatives related to AI governance.

The true test for international cooperation for the rights-based governance of AI, however, will come when someone dares to broaden the discussion from a group of relatively like-minded countries and traditional allies to a truly global forum. While the likelihood of that happening anytime soon seems thin, given the AI-driven "new Cold War", discussing AI governance among a broader and more geographically diverse group of countries certainly has the advantage of enriching the discussion with a wider set of regional perspectives to inform a possible future rapprochement.

REFERENCES:

BEINING, Leonie, Peter Bihr and Stefan Heumann (2020). *Towards a European AI & Society Ecosystem*, Stiftung Neue Verantwortung, February.

BOULANIN, Vincent, Netta Goussac, Laura Bruun and Luke Richards (2020). *Responsible military use of artificial intelligence: Can the European Union Lead the Way in Developing Best Practice?*, Stockholm International Peace Research Institute, November 2020.

BROADBENT, Meredith and Sean Arrieta-Kenna (2021). *AI Regulation: Europe's Latest Proposal is a Wake-Up Call for the United States*, Center for Strategic & International Studies, 18 May 2021.

European Commission (2020). "EU-US: A New Transatlantic Agenda for Global Change", Press Release, 2 December 2020.

European Commission (2021). *Proposal for a Regulation Of The European Parliament And Of The Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts*, 21 April 2021.

European Defence Agency (2020). "Artificial Intelligence in defence", *European Defence Matters*, Issue 19, pp. 34-38.

FRANKE, Ulrike (2021). *Artificial divide: How Europe and America could clash over AI*, European Council of Foreign Relations, 20 January 2021.

GARCIA, Denise (2021). "Stop the emerging AI Cold War", *Nature*, Vol. 593, 13 May 2021.

HEIKKILA, Melissa (2021). "NATO wants to set AI standards. If only its members agreed on the basics", *Politico*, 29 March 2021.

MACCARTHY, Mark and Kenneth Propp (2021). *Machines learn that Brussels writes the rules: The EU's new AI regulation*, Brookings, 4 May 2021.

MUELLER, Benjamin (2021). *Europe's GDPR Regulators' AI Proposals Reveal Their Privacy Fundamentalism*, Center for Data Innovation, 29 July 2021.

National Security Commission on Artificial Intelligence (2021). *Final Report*, March 2021.

NEWMAN, Jessica (2021). "Now is the Time for Transatlantic Cooperation on Artificial Intelligence", *Georgetown Journal of International Affairs*, 13 July 2021.

RAZIS, Evangelos (2021). *Europe's Gamble on AI Regulation*, US Chamber of Commerce, 2 June 2021.

TUCKER, Patrick (2021). "US Needs to Defend Its Artificial Intelligence Better, Says Pentagon No. 2", *Defense One*, 22 May 2021.

Club de Madrid/Boston Global Forum

POLICY LAB
FUNDAMENTAL RIGHTS IN AI & DIGITAL SOCIETIES: TOWARDS AN INTERNATIONAL ACCORD

Sub-committee 3
The elements & process for an international legal framework to protect fundamental rights in AI & digital spaces.

Issues Paper for the workshop on 7-9 September 2021

- What problem are we trying to address?
- Why law? Why AI ethics is not enough
- Why domestic law is not sufficient and why we need a global agreement
- Consensus is possible
- Steps towards a global multilateral Agreement on AI Governance with the aim to ensure governability of States and a stable international order, collective rights and self-determination of humans and their fundamental rights in the age of AI
    - A mandate to explore the opportunity of a global agreement
    - A process of exchange of information on the subject matter, building inter alia on existing processes within UN Agencies
    - Principles for the process of work towards a global agreement
    - Principles on how to delimit the content of the agreement
        - The aim to maintain governability and related questions of control
        - The aim to maintain self-determination of humans
        - The aim to protect global Human Rights
        - A focus on the specific technological risks resulting from the use of self-learning Algorithms and Data
    - Institutional arrangements
        - Mechanisms to create mutual trust
        - Mechanisms of reporting
        - Mechanisms of decision making
        - Mechanisms of dispute settlement and enforcement
        - Relation to pre-existing international law
        - Role of the United Nations Security Council
        - Existing International Law to be inspired by:
            - United nations convention on the law of the sea (UNCLOS)
            - Treaties concluded under the auspices of the International Atomic Energy Agency (IAEA) or related to its work such as those on nuclear safety, liability, non-proliferation
        - Sources of Inspiration in terms of ongoing work on legally binding AI governance:
            - Work in the Council of Europe on a Convention on AI

- Work in the European Union on an AI Act
- Chines work on AI Regulation

What problem are we trying to address?

The opportunities and risks of Artificial Intelligence (AI) seem limitless. While some believe that many of the most challenging problems of this world can be solved through or with the help of AI, other, such as Bill Gates, consider this technology also a risk alike to atomic power and atomic weapons.

There are a number of reasons why AI poses a global risk: First, AI is analysing data through Algorithms, which are supposed to learn, and thus improve their performance, beyond capabilities of humans. While today such performance of AI beyond capabilities of humans is normally limited to very specific tasks, there is a trend to the broadening of these tasks. It cannot be excluded anymore that a general AI is being developed, which surpasses all aspects of human intelligence and eventually thus could dominate humankind in all respects. The problem of control of AI in order to ensure that humans do not become objects of machine control and the ability of states to govern is not undermined has already been amply described in science. It is clear that technology alone cannot deliver such control to a sufficient extent.[1]

Second, much of Artificial Intelligence is and will be delivered and deployed via the Internet, across the borders of this world. It is thus a technology crossing borders in the virtual space without effective controls.

Third, AI is a technology, which is being developed by global corporations and states for various purposes and not limited to a sector. It is a multi-purpose technology, which has potential to scale globally in most areas of our lives, ranging from education and health via the production and delivery of media content and opinions, important for democracy, right through to the management of all essential infrastructures and military purposes. It is the sensitivity of the multiple contexts within which AI is deployed which requires to give its good functioning and governance highest attention. It is also clear that those who control the functioning of AI in sensitive sectors, including both states and global corporations, will command greatest power in this world.

Fourth, AI can be embedded in autonomous machines, which may cross borders in trade, crime or military operations, thus combining classical physical safety and security risks with the new risks of AI.

Fifth, AI performs often nothing more than large scale and rigorous optimisation. The large scale and global reach of AI programmes for optimisation may create major problems if they create disadvantages for many which are often invisible. For example, if AI optimises for one group, namely the owners of a cooperation, it may at the same time create major detriments to another group, for example workers. The power and scale of such rigorous, and often not visible optimisation, is a key global risk of AI, as it can create huge detriment to large groups of people on an unprecedented scale.

It is important to consider this cross border, global and multipurpose nature of AI in any attempt to assess risks and opportunities of AI. The world has become a community of risk, not only relating to COVID, but also relating to AI. And the scarcest resources is not the one next great idea in terms of a technological solution to a major global challenge, but the scarcest resource relating to AI is the ability

[1] Iyad Rahwan a.o., Superintelligence Cannot be Contained: Lessons from Computability Theory, Journal of Artificial Intelligence Research 70 (2021) 65-76, https://pure.mpg.de/rest/items/item_3020363_8/component/file_3283858/content; see also Stuart Russel, Human Compatible – AI and the problem of control, 2019, with references to reviews and different language versions at https://people.eecs.berkeley.edu/~russell/hc.html.

to agree, both within states and among states, on how to govern this new powerful and globally scaled multipurpose technology to the benefit of states and mankind.

It is before this background that we are discussing how to give an impetus to the international community to start work on a global multilateral agreement on the governance of AI.

In a world where even Member States of the UN opt for "club"/plurilateral models such as the G-20, and taking account of the reality that only a few countries in the world are actually producing AI products, it will be important to identify the right forum for such work to start. The fact that all states and all people of this world are likely to be affected by an omnipresent, multipurpose technology like AI is an argument to be considered in favour of placing such initial work under the auspices of the UN. In the course of this work, the question of the right forum to make progress will certainly arise repeatedly. One may also ask whether there is a group of countries which could act as catalysts for a global agreement.

Geo-techno politics must be factored in. This includes the growing discontent with China's growing role in UN agencies and on AI innovations and their use to control people and society.

Also, it is generally considered that developing nations align with China's critiques of the prevailing approach to global governance/multilateral governance. The 'G77' is often overlooked in these discussions. While not necessarily major AI producers, they are the market base for which the major producers (US, Europe, China) will be gunning for market share. As they are thus affected, their voice must be heard.

Why law? Why AI ethics is not enough

How a rights based global agreement will fare in the world that is laden with norms, non binding agreements and models of AI Governance outside the scope of state law is a crucial question. There have been over the last decade numerous attempts to create an Ethics of Artificial Intelligence. Corporations developing AI, in order to shoulder their responsibilities and to give orientation to their Engineers, but also states and multilateral organisations, either to prepare or to substitute legislation, have driven these efforts. The number of publications on AI Ethics in the academic field has become hard to follow. There are now more than 80 catalogues of Ethics for AI. Many of their principles overlap in general, but there are also many differences in important details.

Some of the Ethics Codes of Conduct have been accompanied by governance structures within corporations. However, some of these structures have been abandoned again or have been criticized as ineffective. There is also an emerging concern among business about free riders who will not commit to voluntary codes of Ethics, thus putting into question the level playing field of markets and competition.

The ethics community has a moral hazard when it comes to advocating for law as an instrument to govern AI. It is notable however on the other hand that some ethics committees have called for law to be put in place, in order to ensure the democratic legitimacy, the binding nature of the rules and enforceability of AI governance rules against even the most powerful corporations but also the many potential free riders.

Considering the high potentials as well as high risks associated with AI, and the concerns of business relating to a level playing field for AI, in the European Union a consensus has emerged that binding law is necessary, and that the previous AI ethics and governance exercises on the level of the European Union and elsewhere were a good preparation in terms of identifying the challenges which need to be addressed in law. The European Commission has for this reason proposed an AI Act, which is

presently being negotiated by the legislators. Similar considerations have led the broader Council of Europe to start exploratory work towards a binding Convention on AI.

Much of the private self-governance and non binding rules have value in terms of issue spotting and directional orientation. However, these non-binding instruments alone have not delivered the legal certainty and level playing field, which are necessary to both ensure that technology development can thrive and that this progress is to the benefit of humankind, thus in particular that risks arising from the new technologies and related business models are sufficiently addressed and mitigated.

Why domestic law is not sufficient and why we need a global agreement

On the global Level, both the OECD and the G-20 have already recognised the need for common principles on AI Governance. However, the texts adopted in these fora, while giving orientation to corporations, engineers and domestic lawmakers, suffer from the same deficiency as ethics codes: They only have a character of a political appeal, do not carry democratic legitimacy nor binding nature or enforceability. Executives, not legislators, which normally ratify international law obligations, have thus signed them.

Domestic law, which only the European Union is presently negotiating among legislators, will not be able to fully address the global scalability and cross border nature of AI. In order for all states to be certain that the ability to govern will not be undermined through AI being either used with intent for this purpose or getting out of control and thus undermining governability, it is necessary to create rules and structures through which states can support each other in maintaining control over AI, to the benefit of governability and human rights, thus mankind. Only international law which sets out basic substantial principles for this purpose as well as institutions and mechanisms sufficiently developed to be able to deal with the power accruing to hose developing and controlling AI will be able to serve this purpose.

It is important that the great powers of this world as well as small states all sign up to such a global agreement, as AI can be developed and deployed all over the world, with impacts in all other parts of this world. Legitimacy for establishing an international legal framework arises out of the common interest in governability of states, a peaceful international order and giving effect to existing rules of international law in the technical age, in substantive terms, and in procedural terms, as in all international law, from the ratification by legislators.

Considering that much power arising with AI is in the hands of private companies, the question arises whether these companies should be directly bound by a global agreement and how this could be made possible. Rules directly applicable to private parties can arise from international law. The alternative to this is an international agreement in which states take the obligation to enforce certain rules against private actors under their jurisdiction. In *fine* the state enforcement against private parties of any rules agreed will be key in both cases to give effect to binding rules. [2]

Consensus is possible

Since 2019, successive exercises of consensus building on principles for AI have demonstrated that a global consensus is possible. While none of these texts is binding, they show that the international community is learning about the opportunities and risks of AI. And they express a need to increase

[2] https://scholarship.law.georgetown.edu/cgi/viewcontent.cgi?article=1987&context=facpub&httpsredir=1&referer=

precaution, in light of the COVID experience as well as increasing potentialities of AI and related risks studies, while keeping the way open for the development of AI in the best interest of mankind. In 2019 the OECD Council adopted Recommendations on Artificial Intelligence[3], after long negotiations. While the US had initially opposed the adoption of the principles, it eventually agreed, in a remarkable turn of position. On this basis, the G 20 adopted its human – centred AI principles.[4]

In the run up to the EU – US Summit in Brussels on 14 June, a group of academics from the US and the EU produced a Manifesto "In defence of Democracy and the Rule of Law in the age of AI". This manifesto sets out proposals for legal action related to AI and digital technology and thus marks a new emerging consensus across the Atlantic.[5] Academic consensus does not often lead to political consensus. But the manifesto, signed in the meantime by academics also from other continents, is a sign that the unbridgeable gap which existed in the time of President Obama on regulating the digital economy between the EU and the US is being closed. The manifesto followed the invitation to the US by President of the European Commission Ursula Von der Leyen to start work on an AI Agreement: "We want to set a blueprint for regional and global standards aligned with our values: Human rights, and pluralism, inclusion and the protection of privacy", she said when accepting the World Leader for global Peace and Security award from Governor Dukakis at the Boston Global Forum. [6]

More recently, in July 2021, the UNESCO Representatives of Member States agreed on a draft recommendation on AI governance, to be submitted to the General Conference of UNESCO Member States in November 2021 for adoption. UNESCO is now calling for an international regulatory framework to ensure that AI benefits humanity as a hole.[7]

Both China and Russia are Members of UNESCO. It remains to be seen whether on ministerial level they join the UNESCO consensus. There are clear however signs of the Chinese government increasingly understanding the need to regulate the digital economy, for various reasons, which include governability in light of the technological and economical power being concentrated through AI. Also the Russian Prime Minister Putin has taken a stance on the need for rules and limitations on AI.[8] While these are just initial indications, maintaining governability in light of technologies, which may become autonomous to an extent that makes control impossible, may be a common concern of many governments, including China and Russia. Maintaining governability could thus be a starting point for work towards a consensus on a broader rights based global agreement on binding rules for AI.

Next Steps and issues to be addressed on the way to a global agreement on AI

- A mandate to explore the opportunity of a global agreement

  This mandate should be focussed on making more concrete the understanding of states how far they are already becoming a global community of risks, and thus how far their common interest in global and binding governance rules as well as international mechanisms to enforce these rules in practice go.

- A process of exchange of information on the subject matter

[3] https://www.oecd.org/going-digital/ai/principles/
[4] https://www.mofa.go.jp/files/000486596.pdf, point 3.
[5] https://www.aiathens.org/manifesto
[6] https://ec.europa.eu/commission/presscorner/detail/en/speech_20_2402
[7] https://en.unesco.org/artificial-intelligence/ethics
[8] https://voicebot.ai/2020/12/10/russian-president-vladimir-putin-rejects-idea-of-ai-politician-in-interview-with-sberbank-voice-assistant/

The process of exchange of information should go beyond the United Nations SG Roadmap for Digital Cooperation,[9] in which AI is one topic. It should focus on existing and future capabilities of AI and the related risks, in particular the problem of control of AI and the challenge of establishing global governance of AI. The process should be open to states, academia, civil society and business. Its aim is to increase the global understanding of risks related to AI and what global governance rules and movements to implement such rules (e.g. a Global Alliance for Digital Governance) will be necessary.[10]

- Principles for the process of work towards a global agreement

  The work towards a global agreement should be open to all interested stakeholders, but run by the Secretary General and a group of lead States. It should in a first step aim to identify existing principles under international law and international human rights law which may be put in question or suffer in their implementation from AI and what legal rules and mechanisms are necessary to address these challenges.

- An early agreement should be sought on principles on how to delimit the content of the agreement
  - The aim to maintain governability and related questions of control of AI by humans (rather than humans being controlled by AI).
  - The aim to maintain self-determination of humans
  - The aim to protect universal Human Rights aligned with the HR conventions
  - A focus on the specific technological risks resulting from the use of self-learning Algorithms and Data, already identified by professional associations[11] and science[12], as well as a complementary, lateral risk assessment, incorporating (geo)political, economic and sociocultural risks posed by AI.

- An early agreement should be sought that Institutional arrangements must be put in place to ensure compliance with the legal principles of the agreement, such as:
  - Mechanisms to create mutual trust
  - Mechanisms of reporting
  - Mechanisms of decision making among parties
  - Mechanisms of dispute settlement and enforcement
  - Mechanisms to enforce directly against private parties under certain conditions, given their relevance in the field of AI

  The Relation to pre-existing international law and a possible role of the United Nationals Security Council in relation to the enforcement of the AI Agreement should be explored.

[9] https://www.un.org/en/content/digital-cooperation-roadmap/
[10] See on this also the United Nations Centennial Initiative and the volume of reports on "Remaking the World – Toward an Age of Global Enlightenment", 15. July 2021, https://bostonglobalforum.org/publications/remaking-the-world-the-age-of-global-enlightenment-2/ .
[11] https://standards.ieee.org/industry-connections/ec/autonomous-systems.html
[12] https://futureoflife.org/data/documents/research_priorities.pdf?x72900

- Existing International Law to be inspired by in terms of the challenges faced and the substantial as well as institutional solutions found are, among others:
  - United nations convention on the law of the sea (UNCLOS)
  - Treaties concluded under the auspices of the International Atomic Energy Agency (IAEA) or related to its work such as those on nuclear safety, liability, non-proliferation.

- Sources of Inspiration in terms of ongoing work on legally binding AI governance:
  - Work in the Council of Europe on a Convention on AI[13]
  - Work in the European Union on an AI Act[14]
  - Chinese work on AI Regulation[15]

Paul Nemitz, Brussels 1.09. 2021

The author here expresses his personal opinion and not necessarily that of the European Commission.

[13] https://www.coe.int/en/web/artificial-intelligence
[14] https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai
[15] https://www.china-briefing.com/news/artificial-intelligence-china-shenzhen-first-local-ai-regulations-key-areas-coverage/; but see also https://merics.org/en/report/lofty-principles-conflicting-incentives-ai-ethics-and-governance-china.

TOWARD A FRAMEWORK FOR ARTICIAL INTELLIGENCE INTERNATIONAL ACCORD

by Nazli Choucri
Professor of Political Science
Massachusetts Institute of Technology
Member, Boston Global Forum Board of Thinkers

I. EMERGENT GLOBAL CHALLENGES

Advances in information and communication technologies – global Internet, social media,Internet of Things, and a range of related science-driven innovations and generative and emergent technologies – continue to shape a dynamic communication and information ecosystem for which there is no precedent.

These advances are powerful in many ways. Foremost among these in terms of salience, ubiquity, pervasiveness, and expansion in scale and scope is the broad area of artificial intelligence. They have created a new global ecology; yet they remain opaque and must be better understood—an ecology of “knowns” that is evolving in ways that remain largely “unknown.” Especially compelling is the acceleration of Artificial Intelligence – in all its forms –with far-ranging applications shaping a new global ecosystem for which there is no precedent.

This paper presents a brief view of the most pressing challenges, articulates the logic for worldwide agreement to retain the rule of law in the international system, and highlights salientfeatures of an emergent *Framework for Artificial Intelligence International Accord (AIIA*) as an initial response to this critical gap in the system of international rules and regulations.

II. NEW REALITY – NEW "UNKNOWNS"

The term “artificial intelligence” generally refers to the theory, development, and construction ofcomputer systems able to perform tasks that normally require human intelligence—such as visual perception, speech recognition, decision-making, translation between languages, self-driving cars, and so forth. It also spans efforts to establish machine-human brain connectivity in ways that are highly exploratory and whose implications are yet to be articulated to any great extent.

We are at the beginning of a new era, a world of mind-machine convergence. Its

current logic, situated at the frontiers of biological intelligence and machine intelligence, is generally anchoredin past data and has made possible whole new sources and forms of design space. Fully understanding the scale of the AI domain remains elusive. We have seen a shift from executing instructions by humans to replicating humans, outperforming humans, and transcendinghumans.

Almost everyone appreciates that advances in AI have already altered conventional ways of viewing and managing the world around us. We have created new realities for everyone—aswell as new possibilities. Nonetheless, when all is said and done, the "intelligence" that is "artificial" remains devoid of autonomous *consciousness,* empathy, and perhaps select otherhuman features – such as ethics –so fundamental to humanity and the social order.

It goes without saying that, sooner or later, humans will program machines to generate what weconsider consciousness to be. Already we are seeing major efforts and assessments in that direction.

The expansion of Artificial Intelligence is widely recognized to change our lives in ways yet unimagined. This expansion has created a new global ecology, one that remains opaque andpoorly understood.

III. CALL for ACCORD on ARTIFICIAL INTELLIGENCE

The world of artificial intelligence today is framed by a set of unknowns – *known* unknowns and *unknown* unknowns – where technological innovations interact with the potential for the total loss of human control. Especially elusive is the management of embedded insecurities in applications of this new, ubiquitous technology and the imperatives of safety and sustainability.

But without adequate guidelines and useful directives, the undisciplined use of AI poses risks tothe wellbeing of individuals and creates fertile ground for economic, political, social, and criminalexploitation. The international community recognizes the challenges and opportunities, as well as the problems and perils. Several countries have already announced national strategies to promote the proper use and development of AI. Others may be in the process. Different countries may impose different measures, individually or collectively, but for the most part new innovations and novel applications remain largely unregulated.

At the operational level, there are as yet no authoritative modes and methods for reviewing andregulating algorithms. This is yet another "open" space, in the full sense of the word. We are now faced with a critical imperative, namely, to address head-on the policy issues raised by AI advances and to assess, evaluate, and respond effectively. We must engage in serious dialogue – buttressed by tolerance, learning, and mutual understanding – to converge on principles and practices of an agreement among members of the global society on a strategy togenerate and enhance social benefits and wellbeing for all, shared by all.

At the core of this imperative is the need to establish a common understanding of policy

andpractice, anchored in general principles to help maximize the "good" and minimize the "bad" Given prevailing ambiguities and uncertainties, it is not surprising that the international community has not yet fully grasped the full implications of the new “unknowns”.

While individually, as well as jointly, these new capacities transcend the prevailing frameworks for maintaining order – nationally or internationally – on balance, the overall patterns appear notto generate a semblance of order. Some countries have developed national policies for the cyber domain, most notably regarding cybersecurity, we have created new tradeoffs that must be assessed. We must now focus on critical principles and essential supporting practices for thenew and emerging world that we have created. We must also envisage fundamental “best practices” for realities that have yet to emerge.

3.1 Toward a Worldwide Consensus

We must now re-think and consolidate the best practices for human development, recognizing the power and value of the individual and of society. Much yet is to be done.

An added factor is that AI is also becoming a focus for foreign policy and international cooperation. There is a shared view that no country will be able to compete or meet the needsof its citizens without increasing its AI capacity. At the same time, many countries are now engaged in technology leapfrogging. It is no longer expected, nor necessary, to replicate the stages of economic development of the West—one phase at a time.

While the possibilities are varied and diverse, there is also is a clear awareness of the challenges and opportunities, as well as the problems and perils of and many. are seeking waysof managing their approach to AI. At least 20 countries have announced formal strategies to promote the use and development of AI.

No two strategies are alike; however, there are common themes even among countries who focus on different aspects of AI policy. The most common themes addressed include those pertaining to scientific research, talent development, skills formation, public and private collaboration, visualization for innovation, and data standards and regulation, among others.

Transcending the diversity of situations and orientations, there is a solid foundation of shared goals in the international community, buttressed by the by activities of United Nations agenciesto facilitate operational strategies and assist in implementation of objectives. These include a general appreciation of skills, education, and talent development, public and private policy innovation, attention to fairness, transparency, and accountability, ethics and values of inclusion, reliability, security and privacy, science-policy links, standards for regulations and data development, and digital infrastructure.

In sum, all countries are, or will be, going through a common experience of adapting to and managing unknowns. All of these venues are generally framed within an overarching context ofsustainable development. All of this creates an international atmosphere welcoming of an *International Accord on Artificial Intelligence* on a global scale.

3.2 Logic for AI International Accord

There is a long tradition of consensus-based social order founded on cohesion and agreement, and not the use of force nor formal regulation or legislation. It is often a necessary precursor formanaging change and responding to societal needs.

The foundational logic addresses four premises: *What, Why, Why and How?*

What?

An international agreement on AI is about supporting a course of action that is inclusive and equitable. It is designed to focus on relationships among people, governments, and other keyentities in society.

Why?

To articulate prevailing concerns and find common convergence. To frame ways of addressingand managing potential threats – in fair and equitable ways.

Who?

In today's world, framing an international accord for AI must be inclusive of:

- Individuals as citizens and members of a community
- Governments who execute citizen goals
- Corporate and private entities with business rights and responsibilities
- Civil society that transcends the above
- AI innovators and related technologies, and
- Analysts of ethics and responsibility.

None of the above can be “left out.” Each of these constitutes a distinct center of power and influence, and each has rights and responsibilities.

How?

The starting point for such a *Framework* consists of five basic principles to provide solidanchors for Artificial Intelligence International Accord.

1. ***Fairness and justice for all*** The first principle is already agreed upon in the international community as a powerful aspiration. It is the expectation of all entities – private and public – to treat, and be treated, with fairness and justice.

2. ***Responsibility and accountability for policy and decision—private and public*** The second principle recognizes the power of the new global ecology that will increasingly span all entities worldwide—private and public, developing and developed.

3. ***Precautionary principle for innovations and applications*** The third principle is wellestablished internationally. It does not impede innovation, but supports it. It does not push for regulation, but supports initiatives to explore the unknown with care and caution.

4. ***Ethics-in-AI*** Fourth is the principle of ethical integrity—for the present and the future. Different cultures and countries may have different ethical systems, but everyone, everywhere recognizes and adopts some basic ethical precepts. At issue is incorporating the commonalities into a global ethical system for all phases, innovations,and manifestations of artificial intelligence.

Jointly, these four foundations – *What, Why, Who, How* – create powerful foundations for framing and implementing an emergent *Artificial Intelligence International Agreement.*

IV. TOWARD an ARTIFICIAL INTELLIGENCE INTERNATIONAL ACCORD

The *AIIA Draft Framework* recognizes path breaking initiatives – notably the Budapest *Convention on Cybercrime* and the European Union *General Directive* – that signal specific policies to protect the integrity of information and the values that support this integrity. Inaddition, it recognizes the ongoing deliberations in the European Union regarding the future of AI and best means of supporting EU objectives, as well as those of member states.

Then, too, the *Draft Framework* acknowledges the deliberations of the United States NationalCommission on Artificial Intelligence, and the Report of its results. Consistent with the legal principle of *a rules-based international community*, the Draft *Framework* consists of several initial procedural and operational strategies, as follows:

1. ***Preamble*** to highlight critical values and conditions to help clarify the underlying commonalities among all signatory entities supporting an *AIIA* of worldwide scale and scope,

2. ***Framework Design*** to define the parameters of structure and process for further global deliberation and refinement,

3. ***Operational Measures*** to buttress pragmatic as well as aspirational purposes, and

4. ***Support System*** for realizing, formalizing and implementing an International as well as Global and International Accord on Artificial Intelligence.

Each calls for some articulation.

4.1 Preamble

The Preamble to the *AIIA Framework* is predicated on critical premises that reflect important features of the results-based system that defines today's international community, and are assumed to be operative at the drafting of the Framework. These are assumptions that enableframing of further order, and are stated as follows:

- *Recognizing* accelerated innovations in and applications of AI in diverse facets of the human condition. All advances and applications thereof must be coupled with, and adhere to, the internationally recognized *precautionary principle.*
- *Supporting* the international community's commitment to *human rights.* The potential harms on society inflicted by unrestrained uses of AI must be prevented in all contextsand situations everywhere.
- *Convinced* of the salience of *rights,* commensurate attention must be given to responsibilities.
- *Understanding* the differences and discrepancies among countries in computationalskills and innovations in AI, a worldwide AI educational initiative must be designed toenable "full recognition" of all challenges surrounding AI.
- *Respecting* the diversity of the international community, all parties, public and private, all measures for implementation will be taken by *national* authorities.
- *Acknowledging* that that the dearth of guidelines may evolve into chaotic and undisciplined conditions that undermine benefits of AI to society by enabling exploitationand damage.

4.2 Framework Design

Consistent with the principles the provisions of the Budapest Convention on Cybercrime as wellas the EU General Directives, and respecting the *Social Contract for the AI Age,* the AIIA draft framework is conceived and designed as:

- A multi-stakeholder, consensus-based international agreement to establish common policy and practice in development, use, implementation and applications of AI.
- Anchored in the balance of influence and responsibility among governments, businesses, civil society, individuals, and other entities.
- Respectful of national authority and international commitments and requires assurancesof rights and responsibilities for all participants and decision-entities.

To consolidate the design into a formal International Accord, it is essential to:

- **Review** legal frameworks for AI at various levels of aggregation to identify elements essential for an international AI legal framework,

- **Recognize** methods to prevent abuses by governments and businesses in uses of AI,Data, Digital Technology, and Cyberspace (including attacks on companies, organizations, and individuals, and other venues of the Internet),
- **Consolidate** working norms to manage all aspects of AI innovations, and
- **Construct** and enable response-systems for violations of rights and responsibilities associated with the development, design, applications, or implementation of AI.

V. PROCESS and ESSENTIAL MEASURES

Given that "unrestricted use" of AI is not deemed acceptable by the international community,and a "total ban" may be unreasonable at this point, the *Draft Framework for AIIA* puts forth aset of measures for immediate review, assessment, refinement, and adoption by the international community. These measures are for all relevant actors and entities.

(1) Individual Rights and Responsibilities

The scope of *rights* includes:

- Rights pertaining to Data and the Internet
- Rights to digital and AI related education
- Rights to political participation in AI policy deliberations
- Right to avoid digital damages

And with rights, come *responsibilities* to:

- Avoid digital damages
- Contribute to the common good
- Participate in codes of digital ethics
- Remain cognizant of AI applications
- Refrain from use of malware or distribution of misinformation

(2) Imperatives for National Policy

Governments are required to:

- **Implement** the AI governance policies, standards, and norms adopted by the international community

- **Provide** education for all citizens – "real" or online – with advanced AI technology
- **Design** incentives and directives for responsible AI use
- **Protec**t intellectual property rights without undermining free access to the information commons

(3) Collaboration among States

International collaboration is required to:

- **Support** shared AI policies and common goals
- **Enable** international measures by creating national policies and instruments
- **Reinforce** protection of human rights in AI innovations and uses
- **Develop** common principles and methods to contain and combat misinformation
- **Recognize** the *Social Contract for the AI Age*
- **Establish** a *Worldwide Alliance toward Digital Governance*.

(4) United Nations and International Organizations

These entities are expected to:

- **Enable** and support sustained data collection and analysis
- **Provide** guidelines for worldwide AI knowledge and education
- **Create** support systems for global digital ethics
- **Establish** international legal foundations for management of AI
- **Convene** all willing entities to participate in the framing and forging of international judicial systems devoted to AI applications
- **Contribute** to the United Nations Centennial, notably a Global Enlightenment Prize and international Lecture
- **Reinforce** the AIWS City initiatives to develop and evaluate a People Centered-Economy

(5) Business and Industry

National and international businesses are expected to:

- **Enable** independent audits for transparency, fairness, accountability, and cybersecurit

- **Adopt** common AI values, standards, norms, and data ownership rules with penalties for noncompliance
- **Collaborate** with governments, civil society, and international organizations to help create a people-centered AI, data, and Internet ecosystem
- **Support** sanction-systems to enforce a rules-based international order

(6) Civil Society

These entities are expected to:

- **Monitoring** governments and firms in support of common values and standards
- **Enabling** all forms of voluntary data, analytics and other cooperatives, including the pooling by individuals of their personal data for the benefit of the group or community, conforming to international norms
- **Supporting** an intelligent, thoughtful development and use of knowledge, as well as institutional opportunities for knowledge.

VI. THE SUPPORT SYSTEM for AIIA FRAMEWORK

Based on the internationally recognized *Precautionary Principle*, the support system for AIIA Framework is expected to facilitate and formalize the Framework and its implementation. Thesupports include the following products and processes:

- **Code of Ethics** for AI Developers and AI Users
- **Operational** systems to monitor AI performance by governments, companies, and individuals
- **Certification** for AI Assistants to enable compliance to new standards
- **Establish** a multidisciplinary scientific committee to provide independent review and assessment of innovations in AI and directives for safe and secure application, consistent with human rights and other obligations
- **Enable a Social Contract for the AI Age** to be supported by United Nations, governments, companies, civil society and the international community
- **Consolidate World Alliance for Digital Governance** as the global authority to enforcethe emergent accord
- **Demonstrate** an initial “proof of concept” with implementation and operations evidencedby the experience and record of the AIWS

- **Establish** a *Network of Democratic Resources*—including democratic governments, companies, institutions, foundations, alliances (such as Global AI Action, World Economic Forum, Global Partnership on AI – GPAI, United Nations Academic Impact,UNESCO, UNDP, and UNEP, among others)
- **Support** the *Network of Democratic Resources* with a *Hub* of founding partners: BGF,Club de Madrid, UN Academic Impact for Centennial Celebration
- **Engage** in worldwide deliberations for consensus on a *bottom-up* and *top-down* construction of a *Global Evaluation System* to assess the ethical issues of AI,review operational applications and implementations in practice, and develop the enabling legal mechanisms
- **Explore** uses of AI in various forms of international relations and global exchange, especially new modes of collaboration and innovations in conflict resolution.

VII. END NOTE: CHALLENGES, OPPORTUNITIES, NEXT STEPS

This *End Note* highlights some salient *challenges,* followed by highlights of opportunities**,** and concludes with a brief word of caution.

7.1 The Challenges

Technology and innovation are growing much faster than the regulatory framework anywhere,and most certainly at the international levels. Of course, we do not want regulations to changeat the level of technological change – that would create chaos; you can imagine why and how.

We can expect innovations in AI to grow much faster than has been the case so far – due in large part to new generations being educated in AI early on. We tend to think that the key players are in the AI arena are companies, governments, and academic researchers. We areoverlooking youth as the growth-asset that will buttress both society and AI in the decades to come. It is foolhardy to ignore what are likely to be the *real* challenges, namely, the scale andscope of (a) unknowns, and (b) unknown "unknowns," and (c) their intended and unintended consequences, individually and collectively.

7.2 The Opportunities

The international community has a long and effective record of framing and reaching agreementin almost all areas of interaction. These are especially powerful in areas of standards, quality controls, certifications and so forth. As a result, we should take stock of what we *do know* aboutwhat works *best* in different areas and domains.

Furthermore, *how* and *why* do we know what works best? These questions are designed to

empower researchers, businesses, government agencies, and international entities – private and public – to address how and why? Then, too, given the known "unknowns," what *should* weknow? We have an opportunity to mine our *own* records for the "best fit" with the properties andconditions surrounding the current Artificial Intelligence dilemma.

Among the major opportunities before us is to inquire: What is the best precedent? Is it nuclear power? Is it climate change? What are other high-risk areas? Usually, we respond to such questions long after the fact. But can we avoid this delay? At this point, we have an opportunityto seriously consider the properties of a global accord in AI before we are faced with a major disaster.

Of high value, for example, is to consider and address the role of ethics in courses on innovations in AI, as well as ethics for all uses and users. So, too, it is important to focus on international law relevant to AI. There are many other high-value issues to consider at this point.The reason is this: The lines of political contention are not yet clearly drawn among potentially conflicting perspectives (or countries). Therefore, now is the opportunity to proceed before theseare consolidated into "lines in the sand."

7.3 The Next Steps

At this point governments do not control AI innovation and/or diffusion. Much of the action takenis not in the public sector. Individuals and non-state groups matter and matter a lot. Constituencies are varied and overlapping. Consensus building is essential for society, not onlyfor governments. Any position taken must be in the interest of everyone. Any initiative cannot be seen to dampen innovation or markets.

At the same time, we know from experience that "punishments"—in their various forms, do not work. We are in a world where large firms in the IT and communication business area are very powerful. Many are larger than most countries. The dilemma becomes: Whom to punish?

The immediate next step is establishing a multi-stakeholder support base. This is a necessarystep to get to the point of articulation of interests and negotiations for "best outcomes."

We are now dealing with 21st century realities wherein state coalition building is essential. We could even initiate a global competition among young minds for creating the best internationalagreement on artificial intelligence.